Contents

Context

Given a set of customer data with the following variables, we are asked to build a logistic regression model to predict customer who are potential churners and to interpret the results.

Variable	Description	Type
Churn	1 if customer cancelled service, 0 if not	Categorical
AccountWeeks	number of weeks customer has had active account	Continuous
ContractRenewal	1 if customer recently renewed contract, 0 if not	Categorical
DataPlan	1 if customer has data plan, 0 if not	Categorical
DataUsage	gigabytes of monthly data usage	Continuous
CustServCalls	number of calls into customer service	Continuous
DayMins	average daytime minutes per month	Continuous
DayCalls	average number of daytime calls	Continuous
MonthlyCharge	average monthly bill	Continuous
OverageFee	largest overage fee in last 12 months	Continuous
RoamMins	average number of roaming minutes	Continuous

Business Report

Based on the data set provided and the detailed analysis we have done, the organization should focus on targeted marketing targeting the customers predicted to churn by our best predictive model Model 3 (Described in subsequent sections) to have a positive impact on yearly revenue by 24%. This will also help to retain maximum number of churners while not spending significant amount as retention cost on customers who will stay or leave irrespective of the marketing effort.

Key observations

The following kind of customers has higher probability to churn

- Customer calling service center more than once who has not renewed his contract recently
- Customer calling service center for more than one time has a probability to churn.
- Customer calling service center more five time who has a data plan
- Customer having data plan who haven't renewed his contract
- Customer who haven't renewed his contract
- Customer having higher roaming mins
- Customer having higher overage fee
- Customer having higher monthly charge

The below table is the summary of Model 3 which is the best model that we have built with respect to revenue impact based on which the key observations above are arrived at.

Variables	Estimate	value	Odds	Probability
ContractRenewalYes	-1.96E+00	***	0.14	12.4
CustServCalls2	2.23E-01		1.25	55.6
CustServCalls3	1.31E-01		1.14	53.3
CustServCalls4	1.60E+00	.	4.95	83.2
CustServCalls5	1.66E+01		15392745.27	100.0
CustServCalls6	3.37E+00	***	29.07	96.7
CustServCalls7	4.51E+00	***	91.09	98.9
CustServCalls8	1.88E+01		142624544.51	100.0
CustServCalls9	1.42E+01		1504445.24	100.0
MonthlyCharge	6.19E-02		1.06	51.5
OverageFee	8.30E-02		1.09	52.1
RoamMins	9.60E-02	***	1.1	52.4
DataPlanYes:CustServCalls5	3.66E-01		1.44	59.1
DataPlanYes:CustServCalls6	3.85E+00	*	46.99	97.9
ContractRenewalYes:DataPlanYes	-2.68E+00	*	0.07	6.4
ContractRenewalYes:CustServCalls1	4.24E-01		1.53	60.4
ContractRenewalYes:CustServCalls4	7.08E-01		2.03	67.0
ContractRenewalYes:DataPlanYes:CustServCalls1	8.80E-01		2.41	70.7
ContractRenewalYes:DataPlanYes:CustServCalls2	3.19E+00	*	24.21	96.0
ContractRenewalYes:DataPlanYes:CustServCalls3	1.59E+00		4.90	83.0
ContractRenewalYes:DataPlanYes:CustServCalls4	4.83E+00	*	125.35	99.2

Strategy for customer retention

Our model predicts potential churners based on logistic regression model built on the provided dataset.

If the predicted customer by the model:

- Is calling service center more than once who has not renewed his contract recently should be offered a plan/promotion/offer to elude him to renew his contract without the need of further calling the service center.
- Is Customer calling service center for more than one time has to be attended with caution to avoid any further customer interaction with the service center
- Is calling service center more five time who has a data plan should be attended with caution to avoid any further customer interaction with the service center
- Is having data plan who haven't renewed his contract should be contact and offered a plan/promotion/offer to elude him to renew his contract
- Haven't renewed his contract should be contact and offered a plan/promotion/offer to elude him to renew his contract Customer having higher roaming mins
- Having higher roaming mins should be offered plan/promotion/offer to convert the roaming mins to local mins
- Having higher overage fee should be offered a plan to reduce the overage fee

Assumption:

- We are assuming for these above targeted marketing strategies the organization would need $ 56.36 per customer who will be targeted based on our model prediction. We have assumed the marketing expense ($ 56.36) as average monthly charges from the given data set to articulate the benefit of the model.

Based on the above assumption,

- If the organization does not do any retention strategy it will lose $97962 yearly per 1000 customers which is 14% dip in the yearly revenue which is equivalent to the churn rate. Any increase in churn rate will widen the dip in yearly revenue by the proportion of increase in churn rate.
- If the organization does marketing to all the customers it will make additional yearly revenue of $ 41662 per 1000 customers which is 6% increase in yearly revenue.
- If the organization does a targeted marketing to the customers who are predicted as churners by our model, the organization will make additional yearly revenue of $ 62549.30 per 1000 customers which is 10% increase in yearly revenue.

Conclusion: Using Model 3 the organization can make a positive impact of 24% on the yearly revenue as compared to the current state of not doing any customer retention strategy.

For detailed understanding of the analysis and the proposed model please continue reading.

Exploratory Analysis

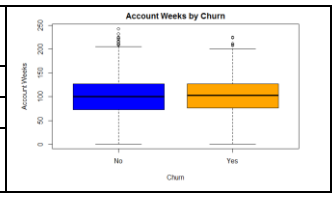
Before building logistic regression model, we looked at the summary information to understand the data that we are dealing with.

Overall Churn rate is 14.49%.

Analyzing Continuous Variable

Account Weeks

Churn	Min	25% of employee	50% of employee	Mean	75% of employee	Max
Overall	1	74	101	101.1	127	243
No	1	73	100	100.8	127	243
Yes	1	76	103	102.7	127	225

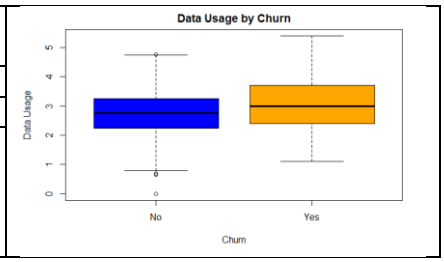


Churning customers have slightly higher mean of account weeks than customers being retained

Data Usage

The following table depicts the data usage for customer with data plan.

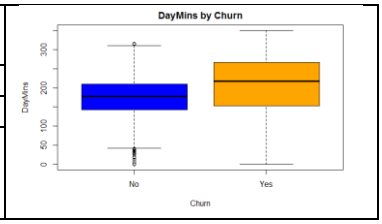
Churn	Min	25% of employee	50% of employee	Mean	75% of employee	Max
Overall	0	2.27	2.78	2.763	3.27	5.4
No	0	2.24	2.75	2.74	3.24	4.75
Yes	1.11	2.417	3	3.01	3.685	5.4



Based on summary data the mean of data usage of churning customer is higher than the customer being retained.

Day Mins

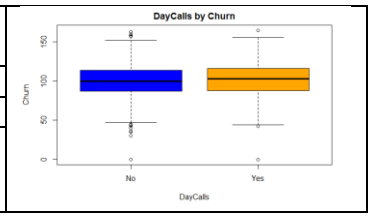
Churn	Min	25% of employee	50% of employee	Mean	75% of employee	Max
Overall	0	143.7	179.4	179.8	216.4	350.8
No	0	142.8	177.2	175.2	210.3	315.6
Yes	0	153.2	217.6	206.9	265.9	350.8



Based on summary data the mean of day mins of churning customers is higher than the customer being retained

Day Calls

Churn	Min	25% of employee	50% of employee	Mean	75% of employee	Max
Overall	0	87	101	100.4	114	165
No	0	87	100	100.3	114	163
Yes	0	87.5	103	101.3	116.5	165



Based on summary data the mean of day calls of churning customers is slightly higher than the customer being retained

Monthly Charge

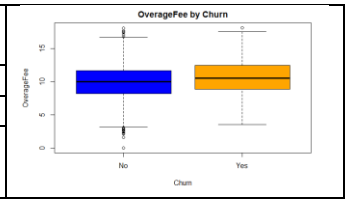
Churn	Min	25% of employee	50% of employee	Mean	75% of employee	Max

Overall	14	45	53.5	56.31	66.20	111.3
No	15.7	45	53	55.82	64.67	111.3
Yes	14	45	63	59.19	69.00	110.0

Based on the summary data the mean of Monthly charge is slightly higher for churning customers.

Overage Fee

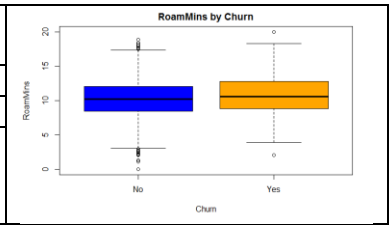
Churn	Min	25% of employee	50% of employee	Mean	75% of employee	Max
Overall	0	8.33	10.07	10.05	11.77	18.19
No	0	8.23	9.98	9.955	11.66	18.09
Yes	3.55	8.86	10.57	10.62	12.47	18.19



Based on the summary data the mean of Overage Fee is slightly higher for churning customers.

Roaming Mins

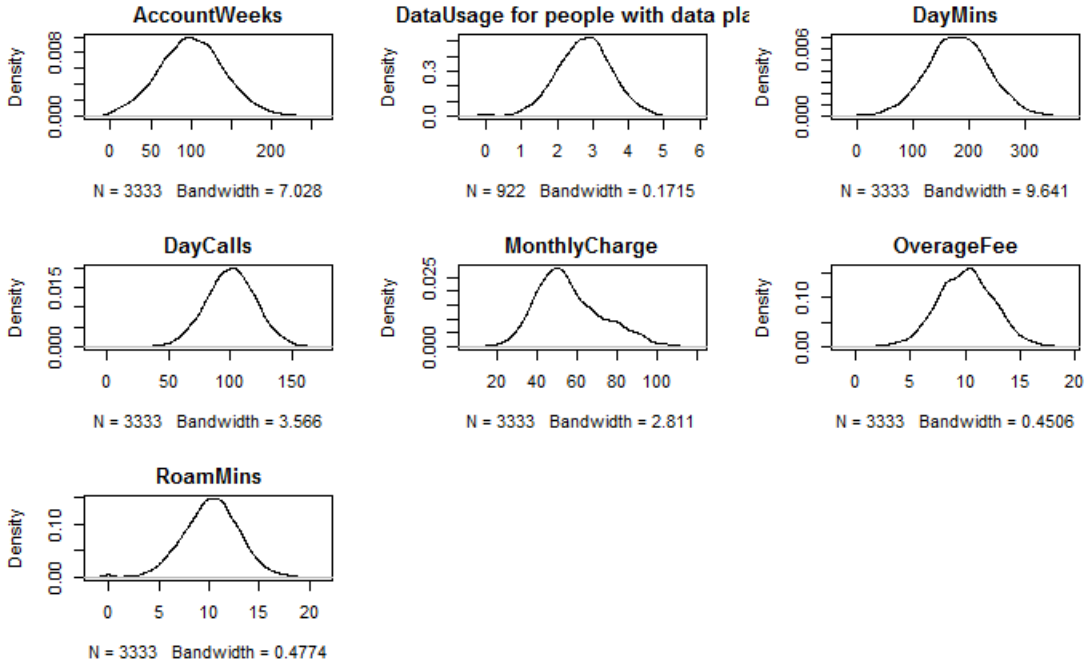
Churn	Min	25% of employee	50% of employee	Mean	75% of employee	Max
Overall	0	8.5	10.3	10.24	12.1	20
No	0	8.4	10.2	10.16	12	18.9
Yes	2	8.8	10.6	10.7	12.8	20



Based on the summary data the mean of Roaming mins is slightly higher for churning customers.

Normality view

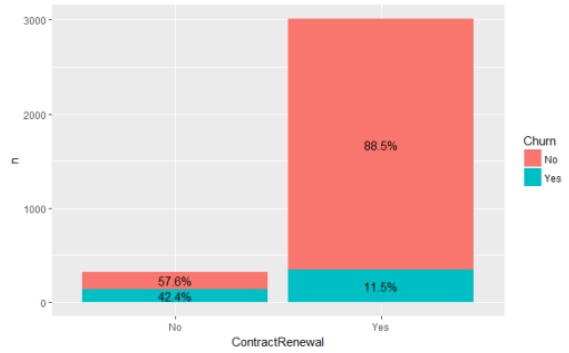
Visibly all the continuous variables seem to be normally distributed.



Note: Data Usage is not considered for customer with no data plan as all such customers will have the data usage as zero which might give a false impressing on the normality of the data.

Analyzing Categorical Variable

Contract Renewal



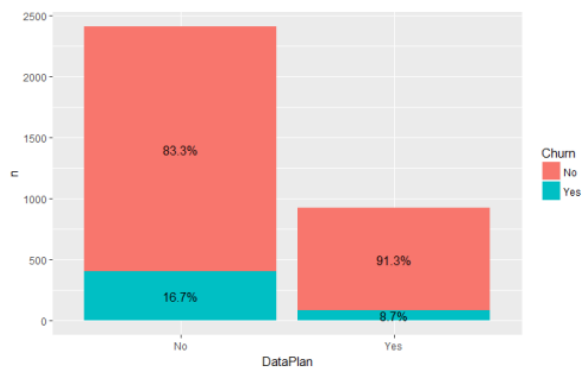
Customers who have renewed their contract recently seem to have churn rate closer to average. Customers who have not renewed their contract recently (may be up for renewal) have significantly higher churn rate than normal

```
        Pearson's Chi-squared test with Yates'
continuity correction

data:  table(data$Churn, data$ContractRenewal)
X-squared = 222.57, df = 1, p-value < 2.2e-16
```

Looking at contract renewal in isolation, there exist a statistically significant relationship between Contract Renewal and attrition.

Data Plan



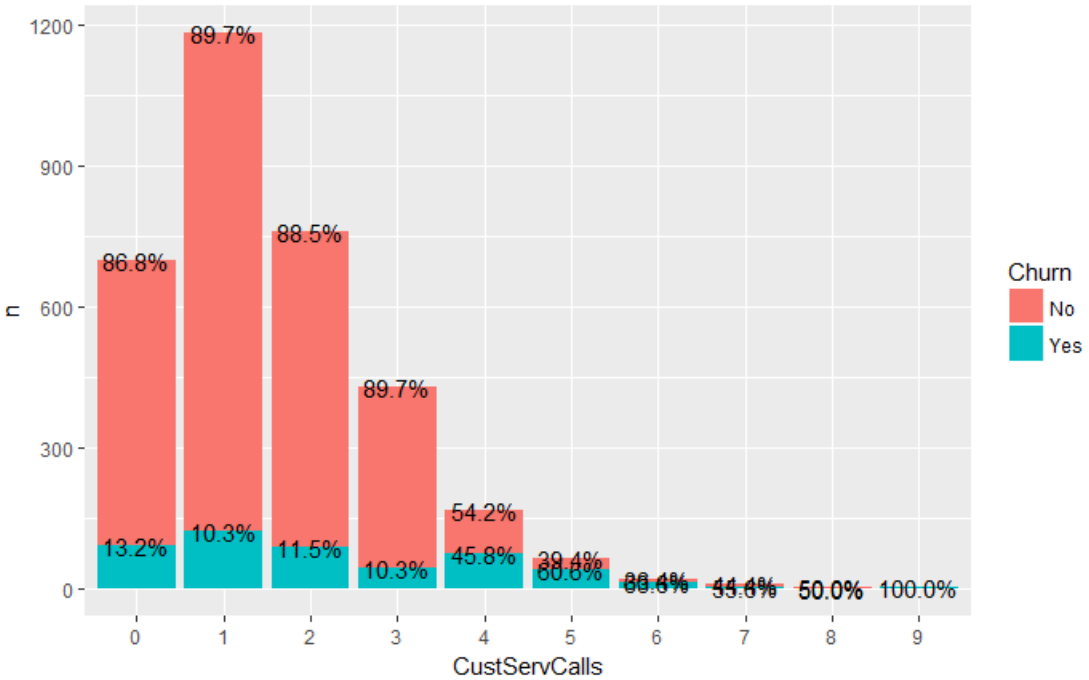
Customers without a data plan are having higher churn rate than average and compared to customers with data plan. Customers with data plan seems to have significantly

```
Pearson's Chi-squared test with Yates' continuity
correction

data:  table(data$Churn, data$DataPlan)
X-squared = 34.132, df = 1, p-value = 5.151e-09
```

Looking ad data plan in isolation, there exist a statistically significant relationship between Data Plan and Churn

Customer Service Calls



Customers who are doing more than 3 calls to service desk seem to have significantly higher churn rate. Customers who are doing less than 4 calls to service desk seem to be having a churn rate around average.

```
Chi-squared approximation may be incorrect
           Pearson's Chi-squared test

data:  table(data$Churn, data$CustServCalls)
X-squared = 342.67, df = 9, p-value < 2.2e-16
```

Since the table values are significantly lower for customers with more than 5 calls to service desk, we cannot rely on this result. We will rely on logistic regression output.

Logistic Regression

Steps

1. Test of Overall Significance of the Logistic Regression
2. Goodness of fit (Pseudo R^2)
3. Individual coefficients and Significance
4. Validate the model
5. Predict
6. Calculate Cut off

Model 1 (Only Main Effect)

We tried GLM on the dev set with Churn as the dependent variable and rest of the variable as the independent variables.

Model

```
logit <- glm(Churn~., data=dev_set, family = binomial)
```

Test of Overall Significance of the Logistic Regression

H_0: Input variables have no significant impact on Response variable (Churn)
H_1: Input variables have significant impact on Response variable (Churn)

```
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 -749.44
2   1 -965.21 -10 431.55  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Since the p-value is overwhelmingly significant we conclude that at 95% confidence Interval we accept the **Alternative Hypothesis** and reject the **Null Hypothesis.**

Goodness of fit (Pseudo R^2)

Based on Pseudo (McFadden) R^2, we conclude that **22.35%** of the uncertainty of the Intercept only model has been explained by the Full Model.

```
         llh      llhNull           G2     McFadden          r2ML          r2CU
-749.4367376 -965.2094900  431.5455047    0.2235502     0.1688745     0.3000436
```

Individual coefficients and Significance

```
Call:
glm(formula = Churn ~ ., family = binomial, data = dev_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0530  -0.5029  -0.3355  -0.1920   3.0586
```

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -6.45852    0.66624  -9.694  < 2e-16 ***
AccountWeeks        0.14298    0.16797   0.851  0.39464
ContractRenewalYes -1.96230    0.17133 -11.453  < 2e-16 ***
DataPlanYes        -1.16051    0.67216  -1.727  0.08425 .
DataUsage          -0.54062    2.33683  -0.231  0.81704
CustServCalls       0.55744    0.04822  11.560  < 2e-16 ***
DayMins             0.16552    2.37565   0.070  0.94445
DayCalls            0.28608    0.33027   0.866  0.38638
MonthlyCharge       0.58114    2.32611   0.250  0.80271
OverageFee          0.06431    0.39652   0.162  0.87116
RoamMins            0.09160    0.02615   3.503  0.00046 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1930.4  on 2332  degrees of freedom
Residual deviance: 1498.9  on 2322  degrees of freedom
AIC: 1520.9

Number of Fisher Scoring iterations: 6
```

Akaike information criterion (AIC): AIC rewards the goodness of fit with an AIC of 1520.9. Contract Renewal (Yes), Customer Service Call and Roam Mins are significant whereas Data Plan (Yes) is marginally significant.

Odds Ratio

In Logistic Regression, the odds Ratio represents the constant effect of a predictor on the likelihood that one unit will occur.

Variables	Odds	Probability
AccountWeeks	1.154	53.6%
ContractRenewal(Yes)	0.141 (Negative)	12.3%
DataPlan(Yes)	0.313 (Negative)	23.9%
DataUsage	0.582 (Negative)	36.8%
CustomerCalls	1.746	63.6%
DayMins	1.180	54.1%
DayCalls	1.331	57.1%
MonthlyCharge	1.788	64.1%
OverageFee	1.066	51.6%
RoamMins	1.096	52.3%

Account Weeks

SIGNIFICANCE:

From the *coefficients table* the predictor ***"AccountWeeks"*** shows a statistically significant *Positive effect* on Response variable ***"Churn"***.

i.e., Increase in AccountsWeek increases the Churn probability.

ODDS RATIO:

For every 100 AccountWeeks increase the probability the customer will churn increases by 53.56% compared to probability of customer not churning, which is 46.44%.

Contract Renewal

SIGNIFICANCE:

From the *coefficients table* the predictor ***"Contract Renewal"*** shows a statistically significant *negative effect* on Response variable ***"Churn"***.
i.e., if a customer has recently renewed the contract there is less probability that he will churn.

Data Plan

SIGNIFICANCE:

From the *coefficients table* the predictor ***"DataPlan"*** shows a statistically marginal negative significance effect on Response variable ***"Churn".***

Customer Service Calls

SIGNIFICANCE:

From the coefficients table the predictor ***"CustSercalls"*** shows a statistically significant Positive effect on Response variable ***"Churn"***.
i.e., increase in Custservcalls will increase the churn probability.

ODDS RATIO

For every 1 CustServcall increase the probability that the customer will churn increases by **63.58%,** compared to probability of customer not churning which is **36.42%.**

Day Mins

SIGNIFICANCE:

From the coefficients table the predictor ***"DayMins"*** shows no significant effect on response variable ***" Churn"***.

ODDS RATIO

For every 1 hour increase the probability that the customer will Churn increase by **54.12%,** compared to probability of customer not churning which is **45.88%.**

Day Calls

SIGNIFICANCE:

From the coefficients table the predictor ***"Daycalls"*** shows no significant effect on response variable **"Ch urn"**.

ODDS RATIO

For every 100 Daycalls increase the probability the customer will churn increase by **57.10%** compared t o probability of customer not churning, which is **42.90%.**

Monthly Charge

SIGNIFICANCE:

From the coefficients table the predictor variable ***"Monthlycharge"*** shows no significant effect on re sponse variable ***"Churn"***.

ODDS RATIO:

For every 10 unit increase in Monthly Charge the probability that the customer will Churn increases by **64.13%** compared to probability of customer not churning, which is **35.87%.**

Overage Fee

SIGNIFICANCE

From the coefficients table the predictor variavle ***"OverageFee"*** shows no significant effect on response variable ***"Churn"***.

ODDS RATIO:

For every unit increase in Overagefee the probability customer will churn increases by **51.60%** c ompared to probability of customer not churning, which is **48.40%.**

Roam Mins

SIGNIFICANCE:

From the coefficients table the input variable ***"RoamMins"*** shows a statistically significant Positive effect on response variable ***"Churn"***.

i.e., increase in RoamMins will increase the Churn probability.

ODDS RATIO:

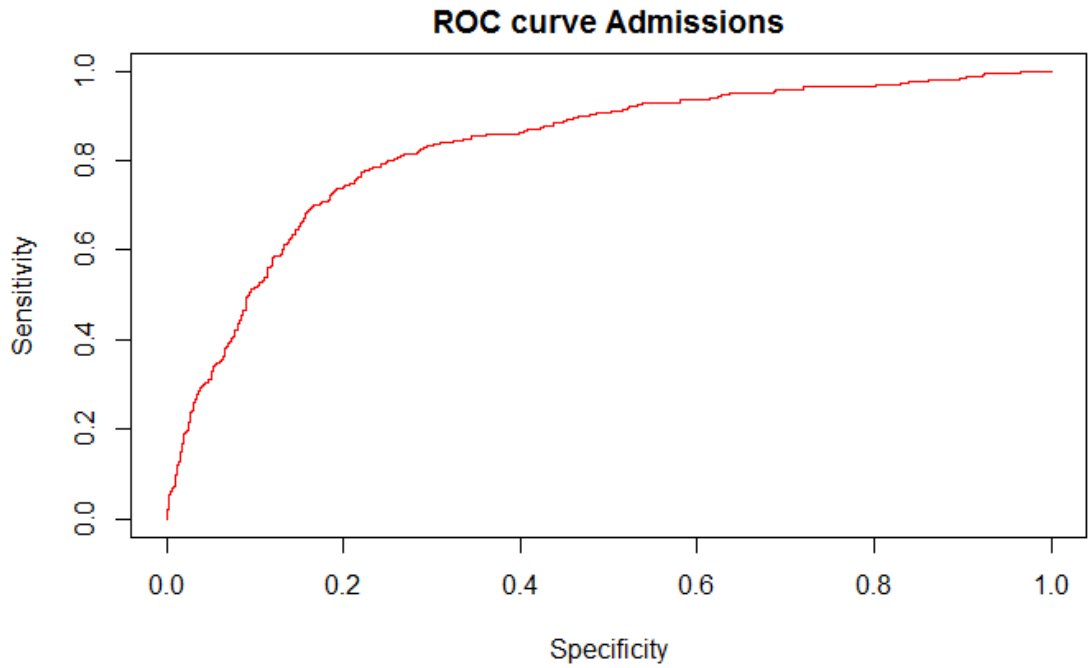
For every unit increase in RoamMins the probability customer will Churn increases by **52.28%** compared to probability of customer not churning, which is **47.72%.**

Model Validation

Confusion Matrix (Dev Set)

	Predicted	
Actual	0	1
No	1946	49
Yes	265	73

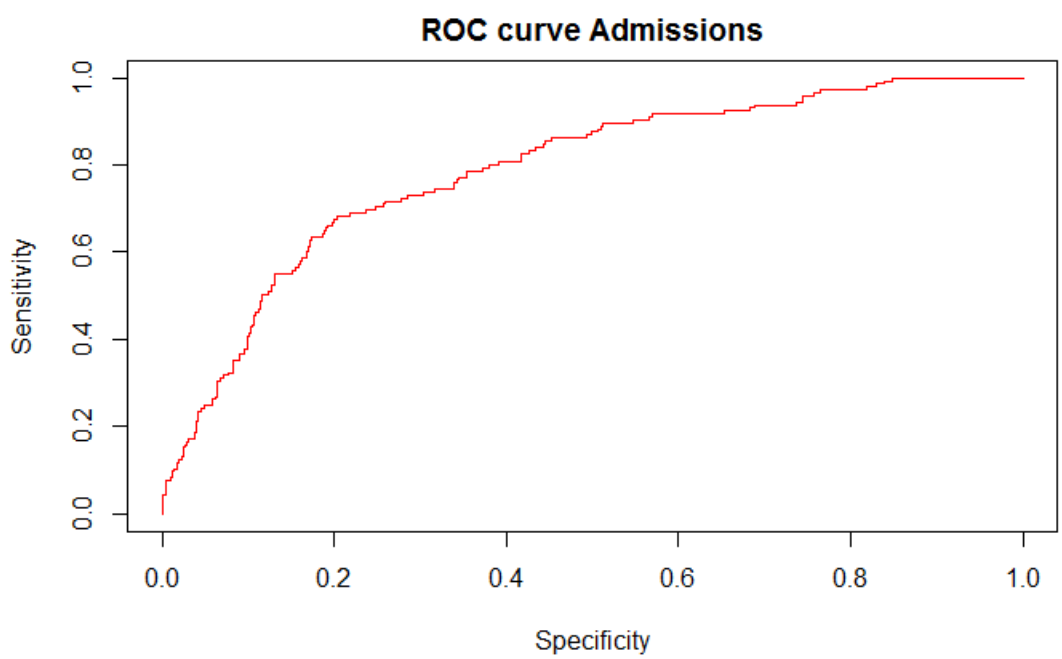
Roc Plot (Dev Set)



Confusion Matrix (Test Set)

```
      predicted
Actual    0   1
   No   832  23
   Yes  123  22
```

Roc Plot (Test Set)



Calculating Cut Off

We have assumed an example of market reach for the group of Customers:

Average marketing Expenses = 56.30 (We have assumed the marketing expense as average of Monthly charges.)
Total earning in a year = 56.3 * 12 = 675.6

P (675.6) + (1-P) (56.30) > 0
On calculating for P, we found the cutoff to be > **0.077**

Confusion Matrix (With Cut Off)

Actual	predicted 0	1
No	451	404
Yes	20	125

If the marketing is done on the predicted model the Profit will be
125 * 667.52 - 529 * 56.30 = 54667.3

Instead if we had to run marketing for the entire population the estimated profit by the Model:
125* 675.6 - 1000*56.30 =28150

Model 2 (Only Main Effect)

Since the customer service calls contains only a limited set of values, this model was tried assuming customer service calls as categorical variable.

Model

```
logit <- glm(Churn~., data=dev_set, family = binomial)
```

Test of Overall Significance of the Logistic Regression

H_0: Input variables have no significant impact on Response variable (Churn)
H_1: Input variables have significant impact on Response variable (Churn)

```
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  19 -706.58
2   1 -965.21 -18 517.25  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Since the p-value is overwhelmingly significant we conclude that at 95% confidence Interval we accept the **Alternative Hypothesis** and reject the **Null Hypothesis.**

Goodness of fit (Pseudo R^2)

Based on Pseudo (McFadden) R^2, we conclude that **26.79%** of the uncertainty of the Intercept only model has been explained by the Full Model. This model has a better McFadden R^2 than Model 1

```
         llh      llhNull           G2     McFadden         r2ML         r2CU
-706.5821353 -965.2094900  517.2547092    0.2679495    0.1988541    0.3533092
```

Individual coefficients and Significance

```
Call:
glm(formula = Churn ~ ., family = binomial, data = dev_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4280  -0.4728  -0.3170  -0.1795   3.1541

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -6.18999    0.69722  -8.878  < 2e-16 ***
AccountWeeks         0.17964    0.17601   1.021 0.307428
ContractRenewalYes  -1.95263    0.17346 -11.257  < 2e-16 ***
DataPlanYes         -1.39471    0.70054  -1.991 0.046492 *
DataUsage           -0.83261    2.42794  -0.343 0.731654
CustServCalls1      -0.08827    0.20065  -0.440 0.660012
CustServCalls2       0.24510    0.21358   1.148 0.251143
CustServCalls3      -0.06333    0.25984  -0.244 0.807431
CustServCalls4       2.18367    0.27735   7.873 3.45e-15 ***
CustServCalls5       3.68220    0.39505   9.321  < 2e-16 ***
CustServCalls6       4.17090    0.62642   6.658 2.77e-11 ***
CustServCalls7       3.33768    0.79388   4.204 2.62e-05 ***
CustServCalls8      16.74384  535.41122   0.031 0.975052
CustServCalls9      13.50530  535.41128   0.025 0.979876
DayMins             -0.08521    2.46683  -0.035 0.972446
DayCalls             0.14826    0.34093   0.435 0.663671
MonthlyCharge        0.90382    2.41535   0.374 0.708257
OverageFee           0.02353    0.41162   0.057 0.954415
RoamMins             0.09558    0.02726   3.507 0.000453 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1930.4  on 2332  degrees of freedom
Residual deviance: 1413.2  on 2314  degrees of freedom
AIC: 1451.2

Number of Fisher Scoring iterations: 12
```

Akaike information criterion (AIC): AIC rewards the goodness of fit with an AIC of 1451.2. Contract Renewal (Yes), Data Plan (Yes), Customer Call (>3), RoamMins has statistically significant impact on the propensity of the customer to churn.

Odds Ratio

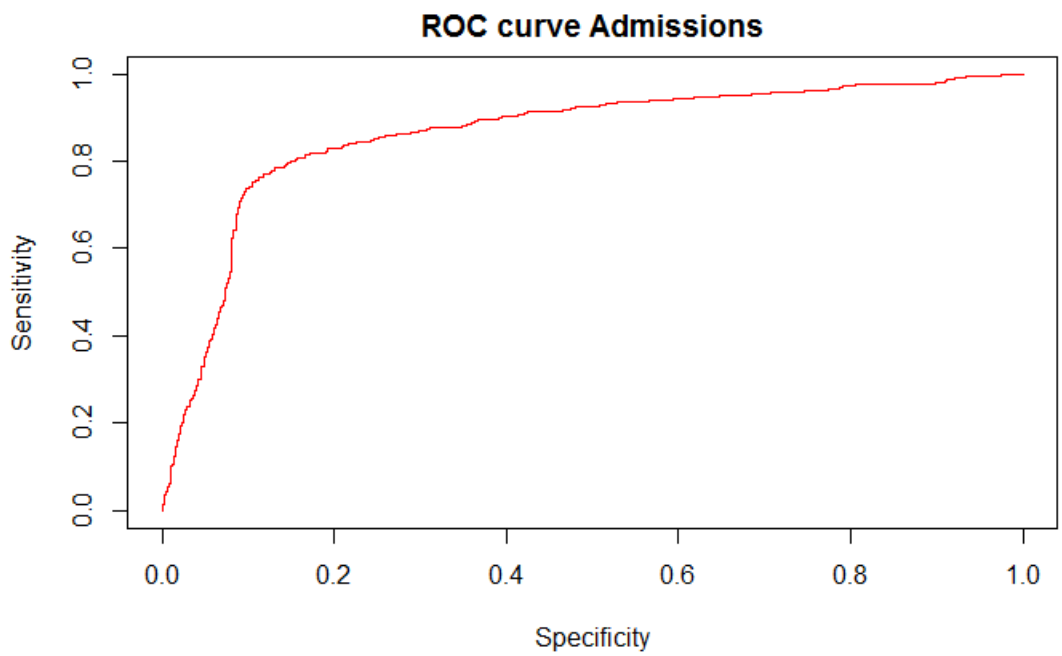
Not discussing the odds ratio as it is discussed in detail in other two models

Model Validation

Confusion Matrix (Dev Set)

```
      Predicted
Actual    0    1
   No  1918   77
   Yes  246   92
```

Roc Plot (Dev Set)



Confusion Matrix (Test Set)

```
      predicted
Actual   0   1
   No  818  37
   Yes 114  31
```

Calculating Cut Off

We have assumed an example of market reach for the group of Customers:

Average marketing Expenses = 56.30 (We have assumed the marketing expense as average of Monthly charges.)
Total earning in a year = 56.3 * 12 = 675.6

P (675.6) + (1-P) (56.30) > 0
On calculating for P, we found the cutoff to be > **0.077**

Confusion Matrix (With Cut Off)

```
      predicted
```

```
Actual   0   1
   No  503 352
   Yes  15 130
```

If the marketing is done on the predicted model the Profit will be
130 * 667.52 - 482 * 56.30 = $59641

Instead if we had to run marketing for the entire population the estimated profit by the Model:
130* 675.6 - 1000*56.30 =31528

Model 3 (With Interaction Effect)

This model considers interaction effect over Model 2.

Before building the model with interaction, we understand the effect of interaction of the categorical variables namely ContractRenewal, DataPlan and CustomerServiceCall.

Note: We tried binning the other continuous variables to understand their interaction, but the model wasn't converging. Need more understanding to perform this.

Model to identify interaction

```
glm(Churn~ContractRenewal+DataPlan+CustServCalls+ContractRenewal*DataPlan*CustServCalls,
data=dev_set, family = binomial)
```

Validity of the model

```
Likelihood ratio test

Model 1: Churn ~ ContractRenewal + DataPlan + CustServCalls + ContractRenewal *
    DataPlan * CustServCalls
Model 2: Churn ~ 1
  #Df  LogLik  Df  Chisq Pr(>Chisq)
1  30 -777.15
2   1 -965.21 -29 376.12  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model is valid and can be used to perform inference.

Model Result

```
Call:
glm(formula = Churn ~ ContractRenewal + DataPlan + CustServCalls +
    ContractRenewal * DataPlan * CustServCalls, family = binomial,
    data = dev_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7941  -0.4679  -0.4618  -0.1665   3.0970

Coefficients: (10 not defined because of singularities)
                                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                                -0.191055   0.310016  -0.616  0.53771
```

```
ContractRenewalYes                          -1.965945   0.363455  -5.409 6.34e-08 ***
DataPlanYes                                 -0.165620   0.582209  -0.284  0.77605
CustServCalls1                              -0.294453   0.404249  -0.728  0.46637
CustServCalls2                              -0.080878   0.454124  -0.178  0.85865
CustServCalls3                              -0.009615   0.546013  -0.018  0.98595
CustServCalls4                               0.884202   0.772081   1.145  0.25212
CustServCalls5                              14.757123 624.193902   0.024  0.98114
CustServCalls6                               2.562465   0.672797   3.809  0.00014 ***
CustServCalls7                               3.543294   1.134015   3.125  0.00178 **
CustServCalls8                              16.723068 882.743396   0.019  0.98489
CustServCalls9                              14.922743 882.743513   0.017  0.98651
ContractRenewalYes:DataPlanYes              -2.464872   1.176133  -2.096  0.03611 *
ContractRenewalYes:CustServCalls1            0.266650   0.469051   0.568  0.56970
ContractRenewalYes:CustServCalls2            0.190185   0.520042   0.366  0.71458
ContractRenewalYes:CustServCalls3           -0.052588   0.624143  -0.084  0.93285
ContractRenewalYes:CustServCalls4            1.183185   0.831818   1.422  0.15491
ContractRenewalYes:CustServCalls5          -11.811666 624.194047  -0.019  0.98490
ContractRenewalYes:CustServCalls6                  NA         NA      NA       NA
ContractRenewalYes:CustServCalls7                  NA         NA      NA       NA
ContractRenewalYes:CustServCalls8                  NA         NA      NA       NA
ContractRenewalYes:CustServCalls9                  NA         NA      NA       NA
DataPlanYes:CustServCalls1                  -0.213870   0.764138  -0.280  0.77957
DataPlanYes:CustServCalls2                  -0.815210   1.044961  -0.780  0.43531
DataPlanYes:CustServCalls3                  -0.732322   1.369058  -0.535  0.59271
DataPlanYes:CustServCalls4                  -1.913822   1.445326  -1.324  0.18545
DataPlanYes:CustServCalls5                   0.165620 805.830975   0.000  0.99984
DataPlanYes:CustServCalls6                   3.611321   1.646508   2.193  0.02828 *
DataPlanYes:CustServCalls7                 -13.321870 624.195664  -0.021  0.98297
DataPlanYes:CustServCalls8                         NA         NA      NA       NA
DataPlanYes:CustServCalls9                         NA         NA      NA       NA
ContractRenewalYes:DataPlanYes:CustServCalls1   0.757138   1.409544   0.537  0.59116
ContractRenewalYes:DataPlanYes:CustServCalls2   2.734792   1.510843   1.810  0.07028 .
ContractRenewalYes:DataPlanYes:CustServCalls3   1.471143   1.997616   0.736  0.46146
ContractRenewalYes:DataPlanYes:CustServCalls4   3.880154   1.837644   2.111  0.03473 *
ContractRenewalYes:DataPlanYes:CustServCalls5   1.165589 805.832044   0.001  0.99885
ContractRenewalYes:DataPlanYes:CustServCalls6         NA         NA      NA       NA
ContractRenewalYes:DataPlanYes:CustServCalls7         NA         NA      NA       NA
ContractRenewalYes:DataPlanYes:CustServCalls8         NA         NA      NA       NA
ContractRenewalYes:DataPlanYes:CustServCalls9         NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1930.4  on 2332  degrees of freedom
Residual deviance: 1554.3  on 2303  degrees of freedom
AIC: 1614.3

Number of Fisher Scoring iterations: 13
```

From the above model, we can find that the following interaction effect exists

- Contract Renewal and Data Plan
- Customer Service Call and Data Plan
- Customer Service Call, Contract Renewal and Data Plan

We built a model with the all the variables (Continuous and Categorical) and the above interaction to understand how it performs.

Interaction Model

```
glm(Churn~.+DataPlan*CustServCalls+ContractRenewal*DataPlan*CustServCalls, data=dev_set, family = binomial)
```

Test of Overall Significance of the Logistic Regression

H_0: Input variables have no significant impact on Response variable (Churn)
H_1: Input variables have significant impact on Response variable (Churn)

```
Likelihood ratio test

Model 1: Churn ~ AccountWeeks + ContractRenewal + DataPlan + DataUsage +
    CustServCalls + DayMins + DayCalls + MonthlyCharge + OverageFee +
    RoamMins + DataPlan * CustServCalls + ContractRenewal * DataPlan *
    CustServCalls
Model 2: Churn ~ 1
  #Df  LogLik  Df  Chisq Pr(>Chisq)
1  37 -688.35
2   1 -965.21 -36 553.73  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Since the p-value is overwhelmingly significant we conclude that at 95% confidence Interval we accept the **Alternative Hypothesis** and reject the **Null Hypothesis.**

Goodness of fit (Pseudo R^2)

Based on Pseudo (McFadden) R^2, we conclude that **28.68%** of the uncertainty of the Intercept only model has been explained by the Full Model. This model has a better McFadden R^2 than Model 1 and Model 2

```
         llh      llhNull           G2     McFadden         r2ML         r2CU
-688.3457249 -965.2094900  553.7275300    0.2868432    0.2112814    0.3753890
```

Individual coefficients and Significance

```
Call:
glm(formula = Churn ~ . + DataPlan * CustServCalls + ContractRenewal *
    DataPlan * CustServCalls, family = binomial, data = dev_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5192  -0.4738  -0.3038  -0.1347   3.4342

Coefficients: (10 not defined because of singularities)
                                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                                 -6.50670    0.75871  -8.576  < 2e-16 ***
AccountWeeks                                 0.22191    0.17850   1.243 0.213796
ContractRenewalYes                          -1.95649    0.39452  -4.959 7.08e-07 ***
DataPlanYes                                 -0.28040    0.96143  -0.292 0.770552
DataUsage                                   -0.52346    2.47170  -0.212 0.832279
CustServCalls1                              -0.38379    0.44560  -0.861 0.389082
CustServCalls2                               0.22300    0.49691   0.449 0.653598
CustServCalls3                               0.13102    0.59677   0.220 0.826225
```

```
CustServCalls4                                  1.59912    0.84649   1.889 0.058878 .
CustServCalls5                                 16.54941  856.71352   0.019 0.984588
CustServCalls6                                  3.36959    0.73048   4.613 3.97e-06 ***
CustServCalls7                                  4.51186    1.14429   3.943 8.05e-05 ***
CustServCalls8                                 18.77573 1455.39756   0.013 0.989707
CustServCalls9                                 14.22393 1455.39766   0.010 0.992202
DayMins                                         0.23235    2.51160   0.093 0.926294
DayCalls                                        0.21294    0.34461   0.618 0.536631
MonthlyCharge                                   0.61890    2.45892   0.252 0.801276
OverageFee                                      0.08297    0.41906   0.198 0.843060
RoamMins                                        0.09597    0.02722   3.526 0.000421 ***
DataPlanYes:CustServCalls1                     -0.37533    0.84144  -0.446 0.655550
DataPlanYes:CustServCalls2                     -1.13121    1.14541  -0.988 0.323345
DataPlanYes:CustServCalls3                     -0.97244    1.43436  -0.678 0.497793
DataPlanYes:CustServCalls4                     -2.95617    1.52702  -1.936 0.052878 .
DataPlanYes:CustServCalls5                      0.36640 1194.57584   0.000 0.999755
DataPlanYes:CustServCalls6                      3.84987    1.67905   2.293 0.021854 *
DataPlanYes:CustServCalls7                    -15.67045 1011.75109  -0.015 0.987643
DataPlanYes:CustServCalls8                           NA         NA      NA       NA
DataPlanYes:CustServCalls9                           NA         NA      NA       NA
ContractRenewalYes:DataPlanYes                 -2.68070    1.21698  -2.203 0.027613 *
ContractRenewalYes:CustServCalls1               0.42422    0.51040   0.831 0.405889
ContractRenewalYes:CustServCalls2              -0.05464    0.56374  -0.097 0.922787
ContractRenewalYes:CustServCalls3              -0.16986    0.67431  -0.252 0.801117
ContractRenewalYes:CustServCalls4               0.70763    0.90827   0.779 0.435922
ContractRenewalYes:CustServCalls5             -12.67073  856.71363  -0.015 0.988200
ContractRenewalYes:CustServCalls6                    NA         NA      NA       NA
ContractRenewalYes:CustServCalls7                    NA         NA      NA       NA
ContractRenewalYes:CustServCalls8                    NA         NA      NA       NA
ContractRenewalYes:CustServCalls9                    NA         NA      NA       NA
ContractRenewalYes:DataPlanYes:CustServCalls1   0.87986    1.46134   0.602 0.547114
ContractRenewalYes:DataPlanYes:CustServCalls2   3.18669    1.59054   2.004 0.045121 *
ContractRenewalYes:DataPlanYes:CustServCalls3   1.58894    2.05333   0.774 0.439029
ContractRenewalYes:DataPlanYes:CustServCalls4   4.83111    1.91762   2.519 0.011758 *
ContractRenewalYes:DataPlanYes:CustServCalls5  -0.01428 1194.57661   0.000 0.999990
ContractRenewalYes:DataPlanYes:CustServCalls6        NA         NA      NA       NA
ContractRenewalYes:DataPlanYes:CustServCalls7        NA         NA      NA       NA
ContractRenewalYes:DataPlanYes:CustServCalls8        NA         NA      NA       NA
ContractRenewalYes:DataPlanYes:CustServCalls9        NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1930.4  on 2332  degrees of freedom
Residual deviance: 1376.7  on 2296  degrees of freedom
AIC: 1450.7

Number of Fisher Scoring iterations: 14
```

Akaike information criterion (AIC): AIC rewards the goodness of fit with an AIC of 1450.7.
Statistically significant main effects:

- ContractRenewal (Yes)
- CustServCalls (>3)
- RoamMins
- DataPlan (Yes)

- CustServCalls (6)

Statistically significant interaction effects:

- CustServCalls (>3) * DataPlan (Yes)
- ContractRenewal (Yes) * DataPlan (Yes)
- ContractRenewal (Yes) * DataPlan (Yes) * CustServCalls (2)
- ContractRenewal (Yes) * DataPlan (Yes) * CustServCalls (4)

Marginal significant interaction effects:

- DataPlan (Yes) * CustServCalls (4)

Odds Ratio

Variables	Estimate	value	Odds	Probability
AccountWeeks	2.22E-03		1	50.1%
ContractRenewalYes	-1.96E+00	***	0.14	12.4
DataPlanYes	-2.80E-01		0.76	43.0
DataUsage	-5.24E-01		0.59	37.2
CustServCalls1	-3.84E-01		0.68	40.5
CustServCalls2	2.23E-01		1.25	55.6
CustServCalls3	1.31E-01		1.14	53.3
CustServCalls4	1.60E+00	.	4.95	83.2
CustServCalls5	1.66E+01		15392745.27	100.0
CustServCalls6	3.37E+00	***	29.07	96.7
CustServCalls7	4.51E+00	***	91.09	98.9
CustServCalls8	1.88E+01		142624544.51	100.0
CustServCalls9	1.42E+01		1504445.24	100.0
DayMins	3.87E-03		1	50.1
DayCalls	2.13E-03		1	50.1
MonthlyCharge	6.19E-02		1.06	51.5
OverageFee	8.30E-02		1.09	52.1
RoamMins	9.60E-02	***	1.1	52.4
DataPlanYes:CustServCalls1	-3.75E-01		0.69	40.7
DataPlanYes:CustServCalls2	-1.13E+00		0.32	24.4
DataPlanYes:CustServCalls3	-9.72E-01		0.38	27.4
DataPlanYes:CustServCalls4	-2.96E+00	.	0.05	4.9
DataPlanYes:CustServCalls5	3.66E-01		1.44	59.1
DataPlanYes:CustServCalls6	3.85E+00	*	46.99	97.9
DataPlanYes:CustServCalls7	-1.57E+01		0.00	0.0
ContractRenewalYes:DataPlanYes	-2.68E+00	*	0.07	6.4
ContractRenewalYes:CustServCalls1	4.24E-01		1.53	60.4
ContractRenewalYes:CustServCalls2	-5.46E-02		0.95	48.6
ContractRenewalYes:CustServCalls3	-1.70E-01		0.84	45.8
ContractRenewalYes:CustServCalls4	7.08E-01		2.03	67.0
ContractRenewalYes:CustServCalls5	-1.27E+01		0.00	0
ContractRenewalYes:DataPlanYes:CustServCalls1	8.80E-01		2.41	70.7
ContractRenewalYes:DataPlanYes:CustServCalls2	3.19E+00	*	24.21	96.0
ContractRenewalYes:DataPlanYes:CustServCalls3	1.59E+00		4.90	83.0

ContractRenewalYes:DataPlanYes:CustServCalls4	4.83E+00	*	125.35	99.2
ContractRenewalYes:DataPlanYes:CustServCalls5	-1.43E-02		0.99	49.6

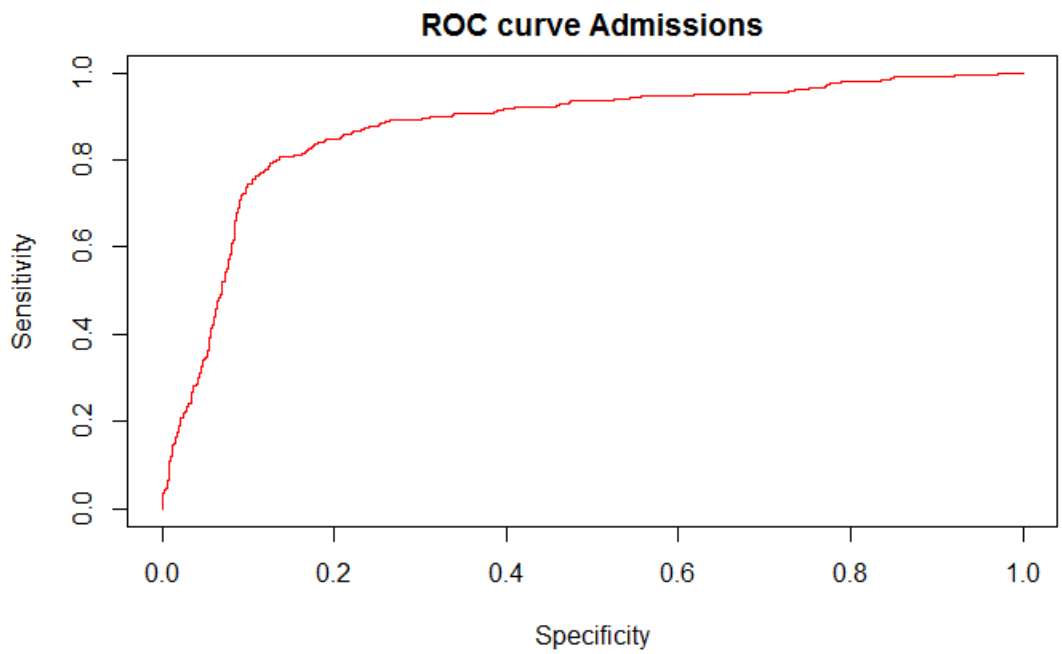
The effects (both main & interaction) marked in green have significant impact on the outcome of churn.

Model Validation

Confusion Matrix (Dev Set)

```
      Predicted
Actual    0    1
   No  1923   72
   Yes  244   94
```

Roc Plot (Dev Set)



Confusion Matrix (Test Set)

```
      predicted
Actual   0   1
   No  818  37
   Yes 114  31
```

Calculating Cut Off

We have assumed an example of market reach for the group of Customers:

Average marketing Expenses = 56.30 (We have assumed the marketing expense as average of Monthly charges.)
Total earning in a year = 56.3 * 12 = 675.6

P (675.6) + (1-P) (56.30) > 0

On calculating for P, we found the cutoff to be > **0.077**

Confusion Matrix (With Cut Off)

```
      predicted
Actual    0   1
   No   503 352
   Yes   12 133
```

If the marketing is done on the predicted model the Profit will be
133 * 675.6 - 485 * 56.30 = $62549.30

Model Comparison

Model	ROC	Test Set (Cutoff = 0.5)			Test Set (Cutoff = 0.077)			Revenue Impact
		Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	
Model 1 (Main)	86.30%	85.4%	87.12%	48.89%	57.6%	95.5%	23.63%	$54667
Model 2 (Main)	86.30%	84.5%	87.23%	41.67%	63.3%	97.10%	26.97%	$59641
Model 3 (Interaction)	87.37%	85%	86.91%	44.44%	63.6%	97.67%	27.42%	$62549

Based on the test data performance with a cutoff of 0.077 Model 3 is the best model. Model 3 performs well in terms of revenue impact.

Reference

The submission includes four RMD files and an R file which has the detailed approach and analysis.

Exploratory analysis: **Exploratory.Rmd** and **draft1.R**

Model 1: **Main.Rmd**

Model 2: **Main_Proper.Rmd**

Model 3: **Interaction.Rmd**