Telecom Churn Case Study

DS C43 - Rajeev Balakrishnan, Abiram Sundar, Karthik G

Background & Problem Definition

Background

In the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another. In this highly competitive market, the telecommunications industry experiences an average of 15-25% annual churn rate. Given the fact that it costs 5-10 times more to acquire a new customer than to retain an existing one, customer retention has now become even more important than customer acquisition.

Actual Problem definition:

To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.

Data Dictionary

Data Dictionary

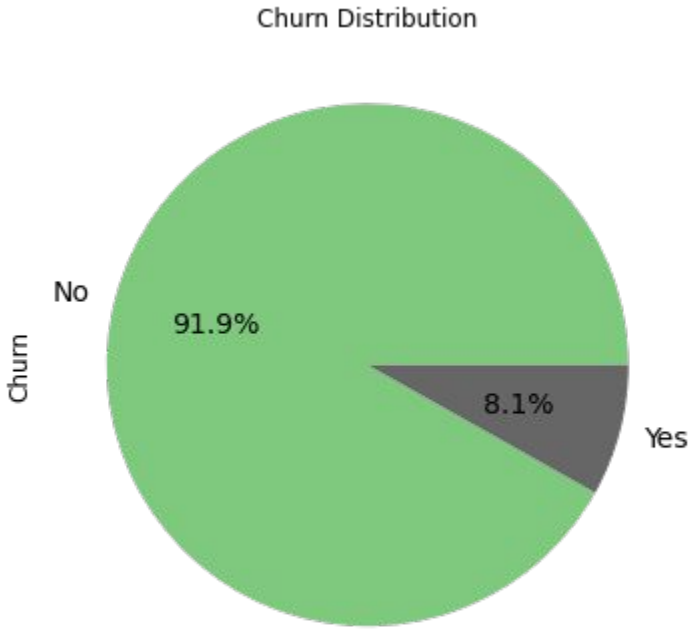
LOC Local calls - within same telecom circle
STD STD calls - outside the calling circle
IC Incoming calls
OG Outgoing calls
T2T Operator T to T, i.e. within same operator (mobile to mobile)
T2M Operator T to other operator mobile
T2O Operator T to other operator fixed line
T2F Operator T to fixed lines of T
T2C Operator T to it's own call center
ARPU Average revenue per user
MOU Minutes of usage - voice calls
AON Age on network - number of days the customer is using the operator T network
ONNET All kind of calls within the same operator network
OFFNET All kind of calls outside the operator T network

ROAM Indicates that customer is in roaming zone during the call
SPL Special calls
ISD ISD calls
RECH Recharge
NUM Number
AMT Amount in local currency
MAX Maximum
DATA Mobile internet
3G 3G network
AV Average
VOL Mobile internet usage volume (in MB)
2G 2G network
PCK Prepaid service schemes called - PACKS
NIGHT Scheme to use during specific night hours only
MONTHLY Service schemes with validity equivalent to a month
SACHET Service schemes with validity smaller than a month
*.6 KPI for the month of June
*.7 KPI for the month of July
*.8 KPI for the month of August
*.9 KPI for the month of September
FB_USER Service scheme to avail services of Facebook and similar social networking sites
VBC Volume based cost - when no specific scheme is not purchased and paid as per usage

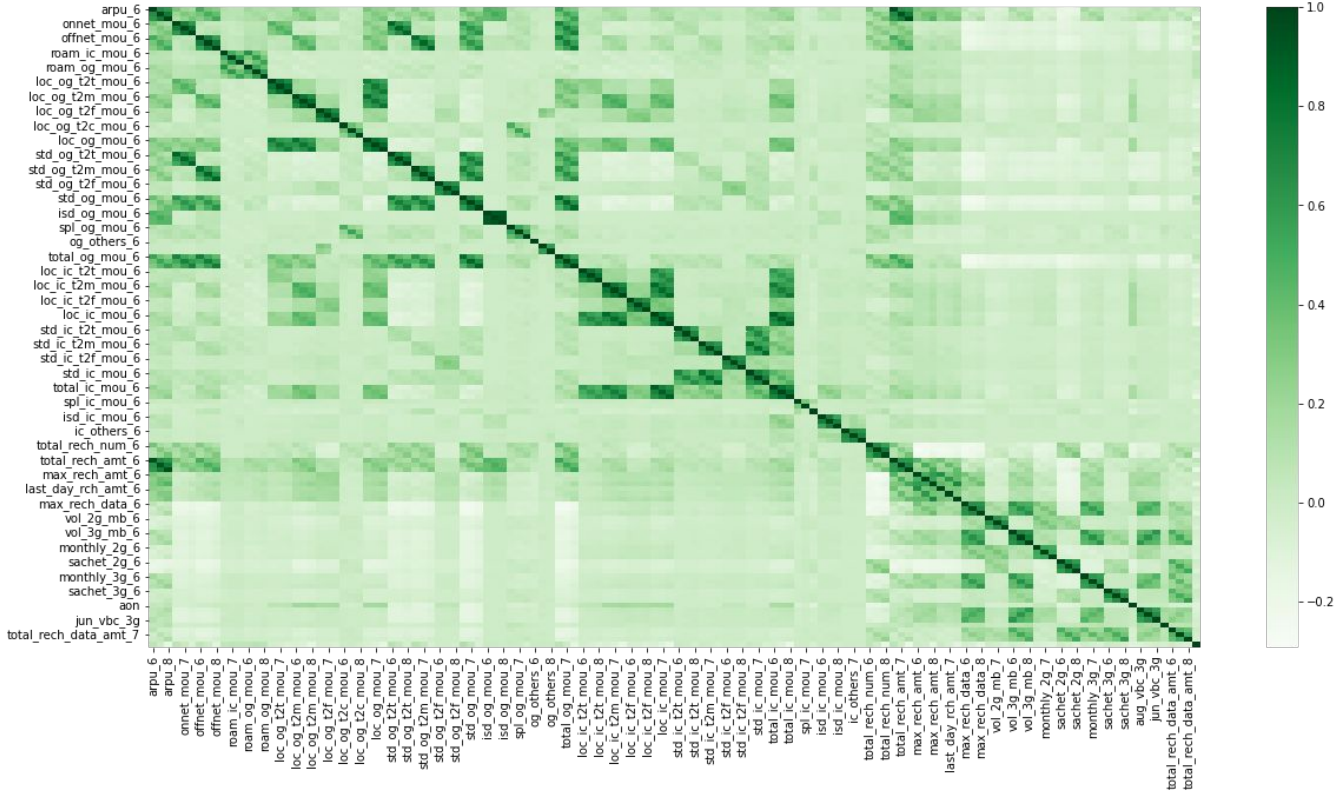
Approach to the Case Study

- Load the Dataset into the Colab Notebook
- Assess & review the structure & headings of the dataset
- Filter High value Customers
- Build the Model
  - using Logistic Regression
  - Using Logistic Regression with balanced class weight
  - Using Logistic Regression with SMOTE (**Synthetic Minority Oversampling Technique)**
  - Decision Tree Classifier
  - Hyperparameter tuning on Decision Tree Classifier
  - Random Forest Classifier
  - Hyperparameter tuning on Random Forest Classifier
  - XGBoost Classifier
  - Hyperparameter tuning on XGBoost Classifier
- Final Score of all models
- Conclusion

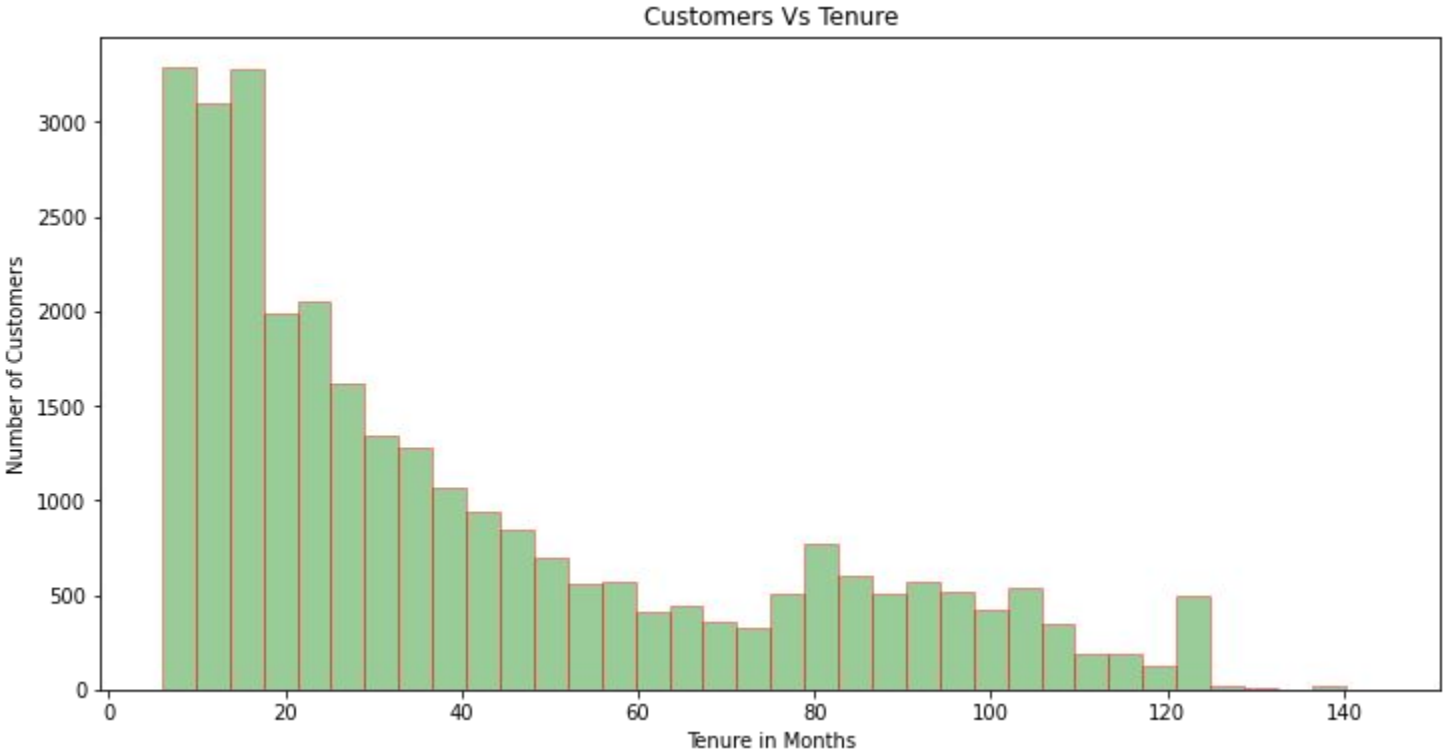
Churn Distribution



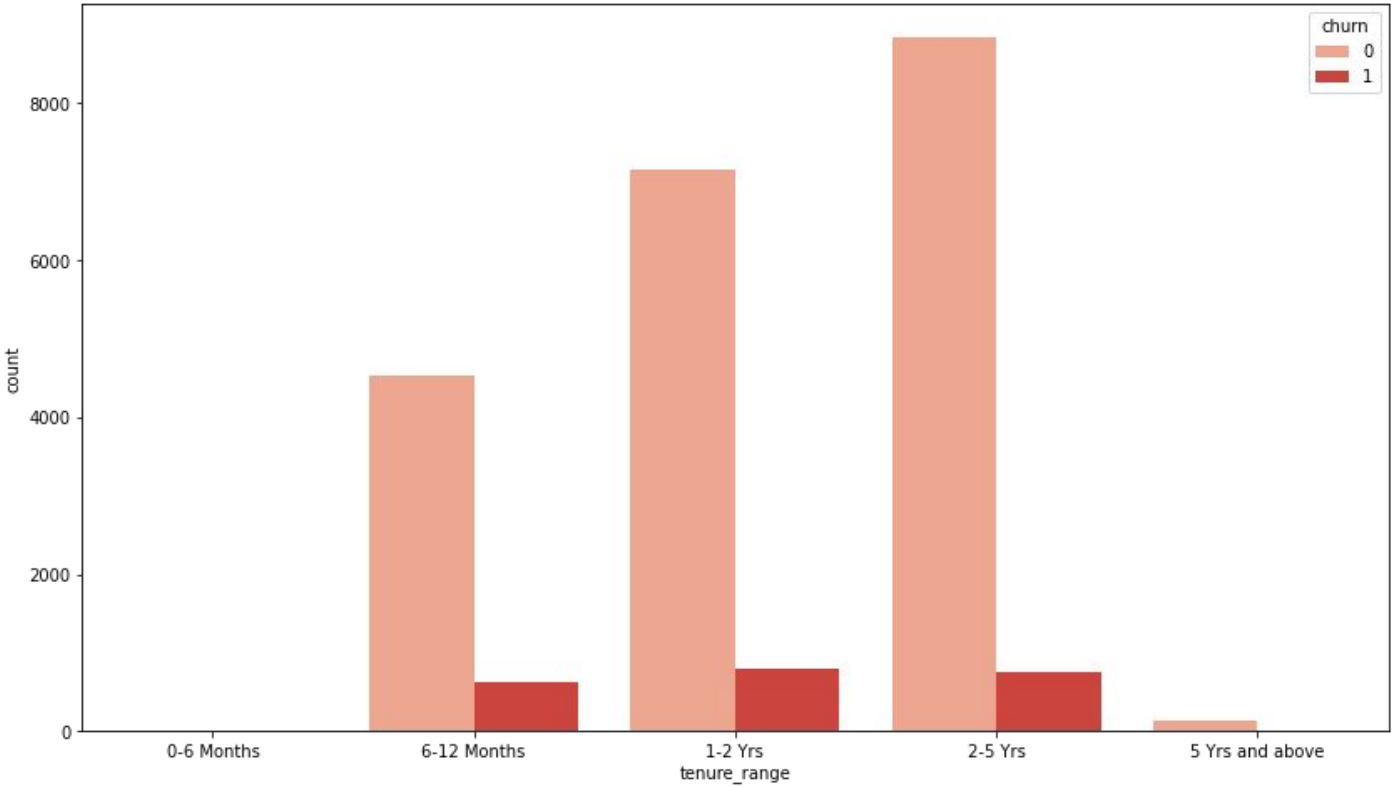
Correlation HeatMap



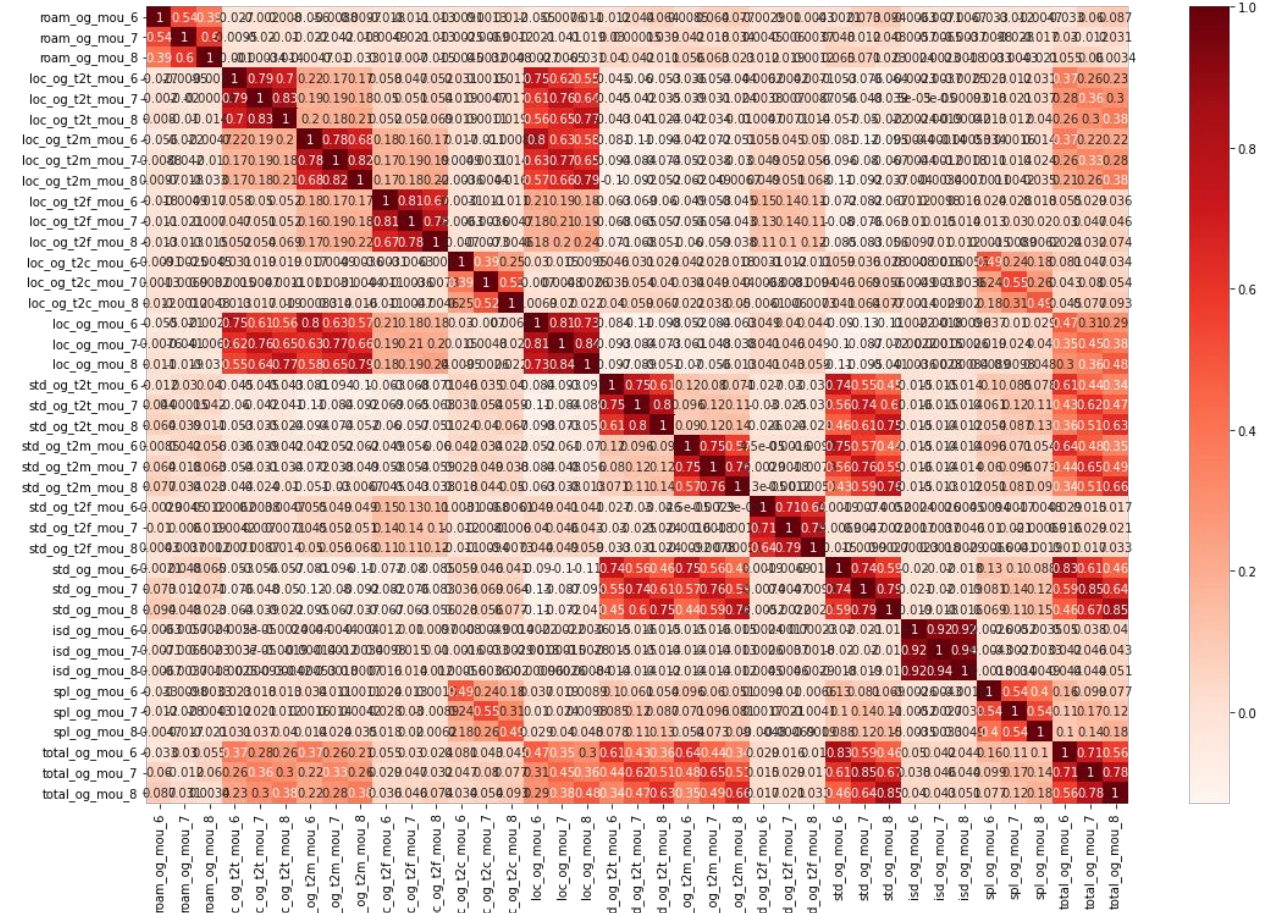
Customer vs Tenure Analysis



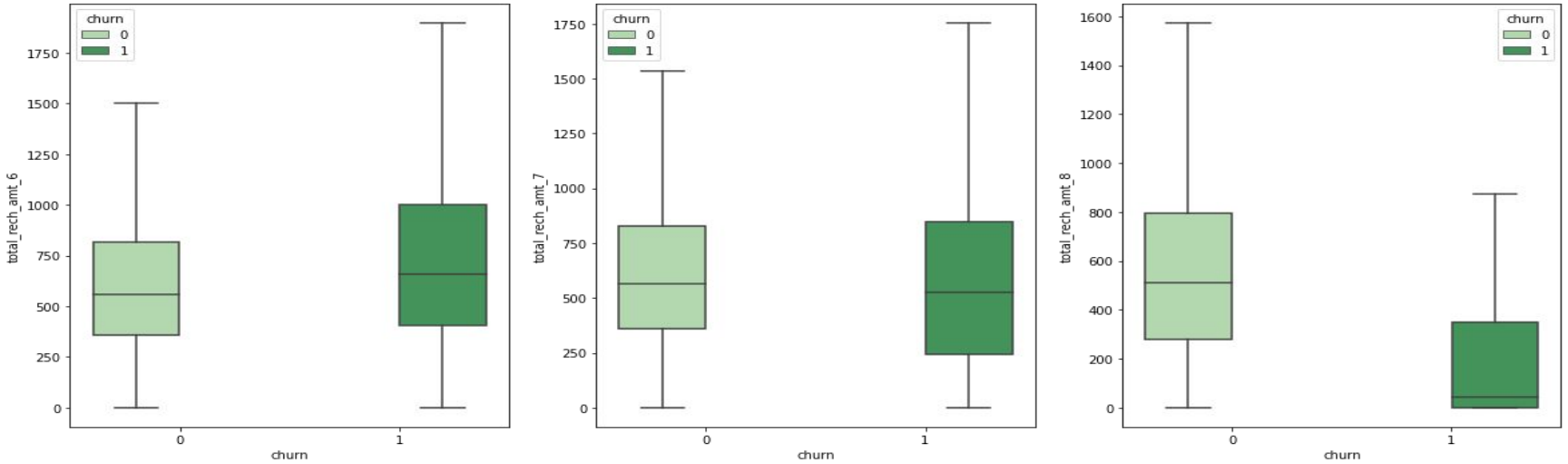
Tenure Binning for Churn & Non Churn Cases



Heatmap for Outgoing Calls

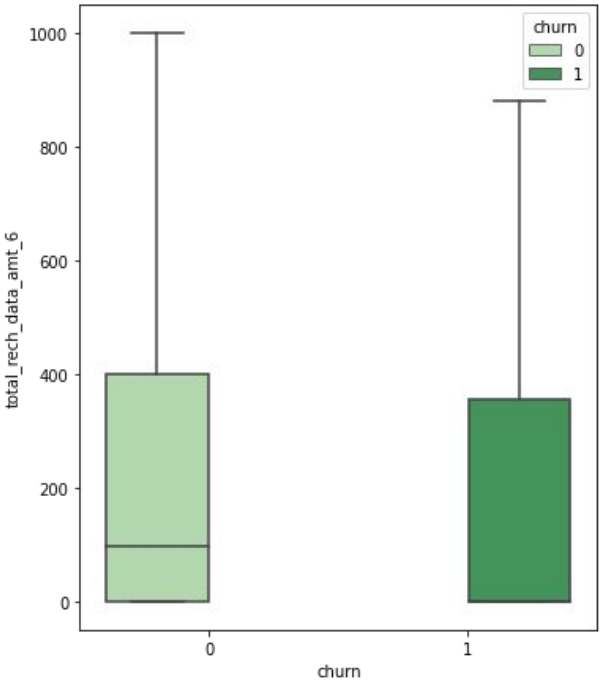


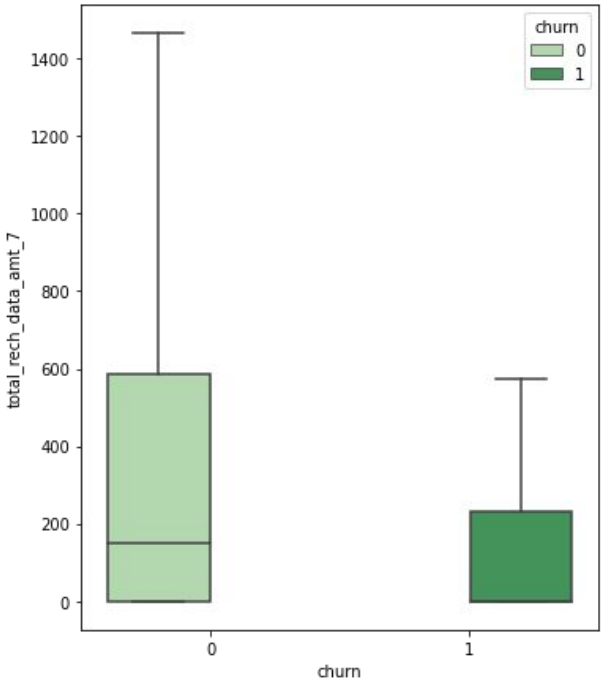
Boxplots for Total Recharge Amount (Month 6,7,8)

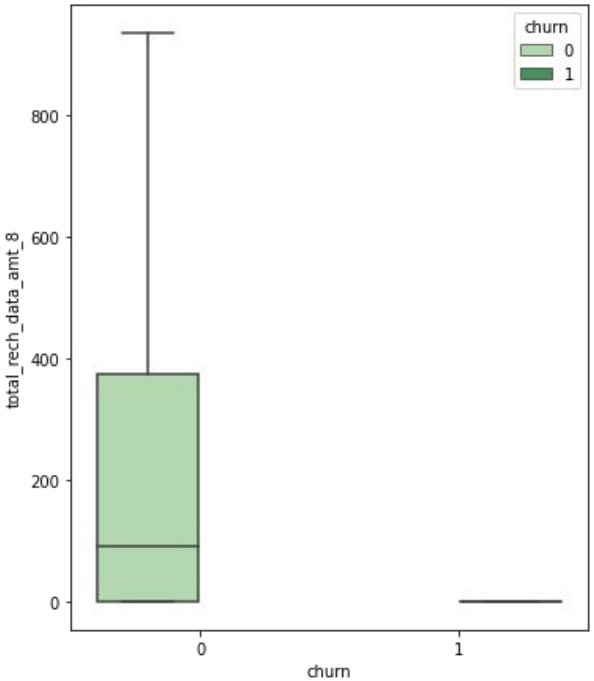


Finding : We can see a drop in the total recharge amount for churned customers in the 8th Month (Action Phase).

Boxplots for Total Recharge Data Amount (Month 6,7,8)

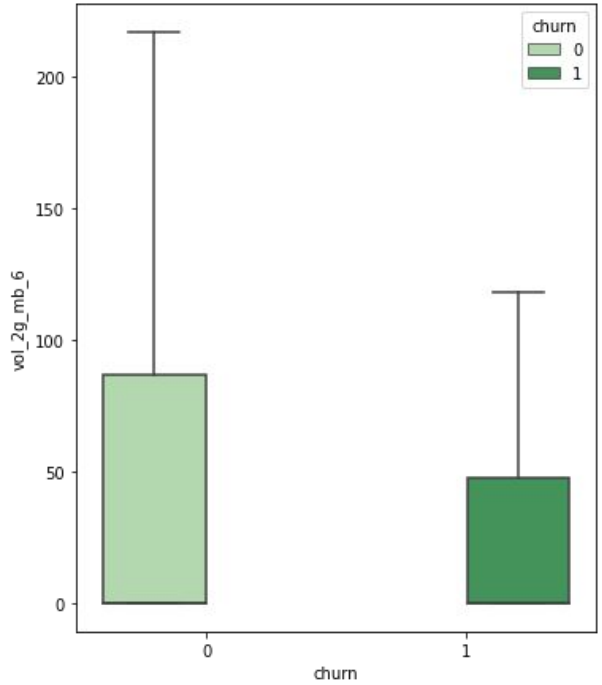


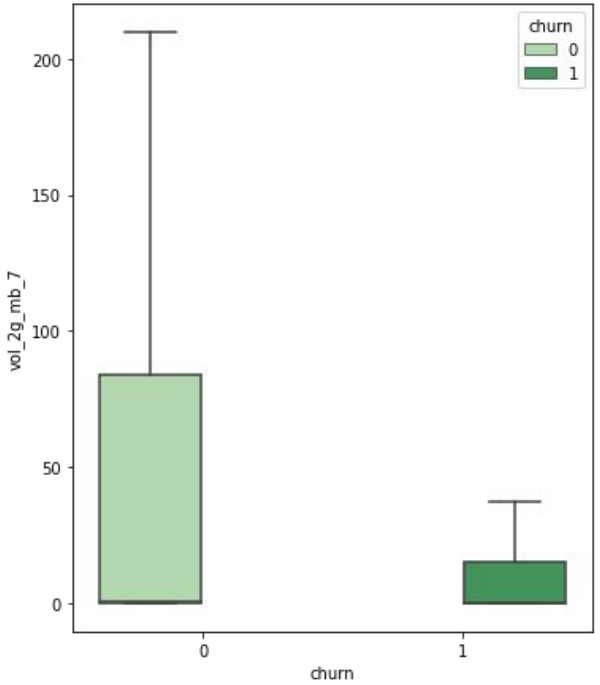


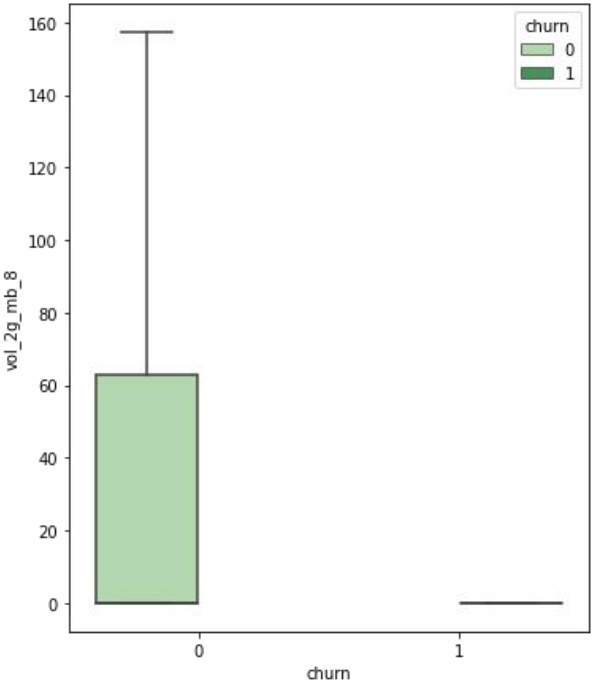


Finding : We can see that there is a huge drop(negligibly less) in total recharge amount for data in the 8th month (action phase) for churned customers.

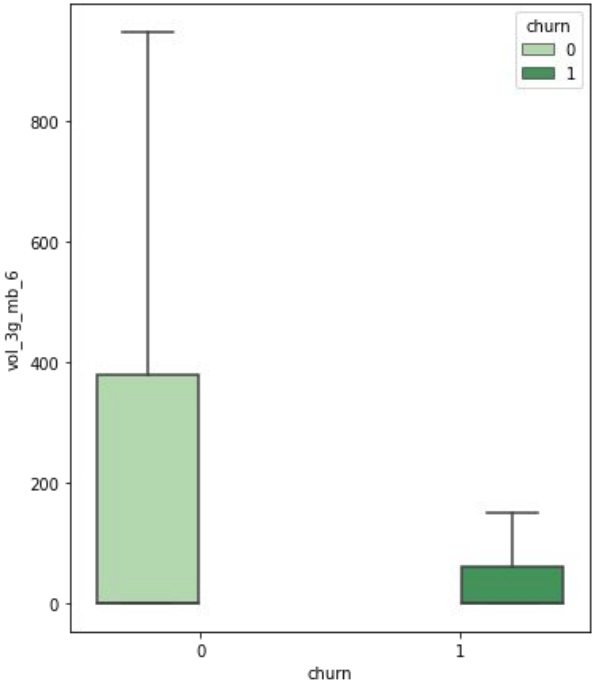
Boxplots for Volume of 2G Usage

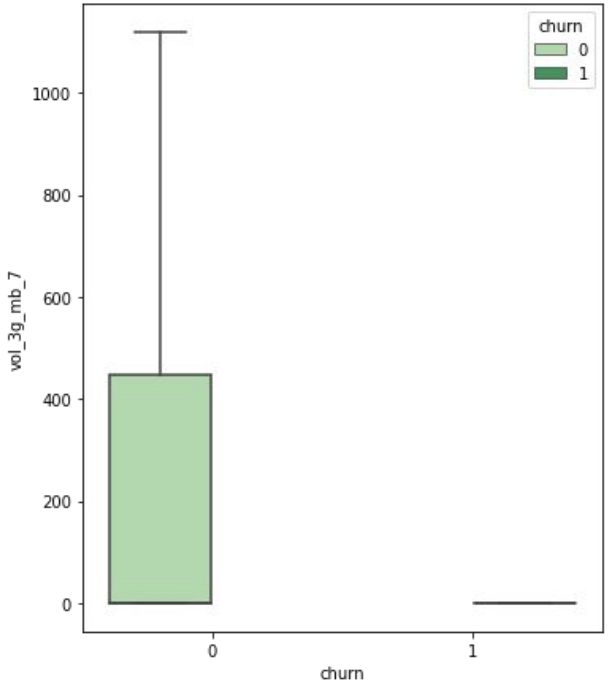


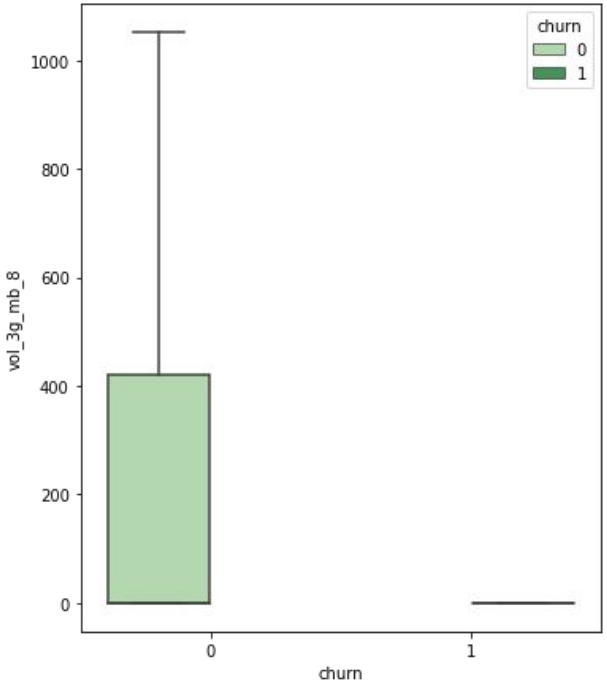




Boxplots for Volume of 3G Usage

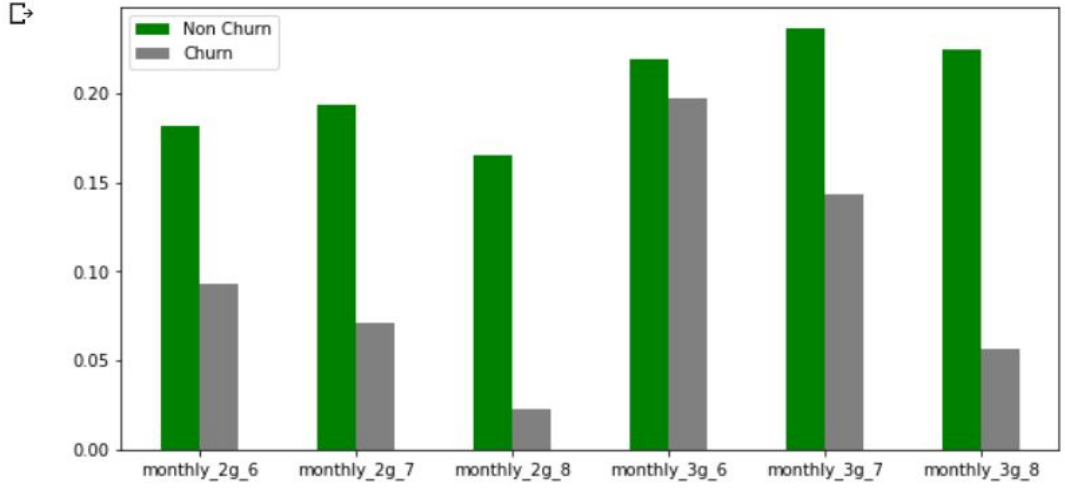






- We have two observations from above: 2G and 3G usage for churned customers dropped in the 8th month
- Also 2G/3G usage is higher for non-churned customers.

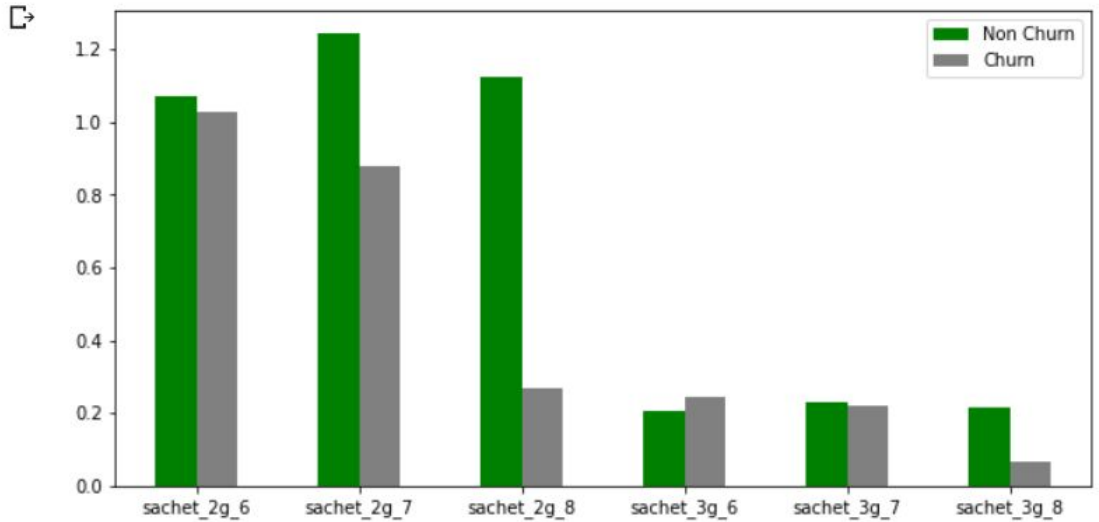
Plot for Monthly Subscriptions for 2G & 3G Usage



	monthly_2g_6	monthly_2g_7	monthly_2g_8	monthly_3g_6	monthly_3g_7	monthly_3g_8
Non Churn	0.181676	0.193215	0.165203	0.219557	0.236575	0.224673
Churn	0.093404	0.070873	0.022941	0.197460	0.143794	0.056944

- Analysis: We can see a drop in monthly subscription for churned customers in the 8th Month

Plot for Smaller Sachets



	sachet_2g_6	sachet_2g_7	sachet_2g_8	sachet_3g_6	sachet_3g_7	sachet_3g_8
Non Churn	1.069303	1.243832	1.124383	0.206313	0.228048	0.214550
Churn	1.029496	0.877509	0.269971	0.244162	0.221221	0.065137

- Analysis : We can see the drop in sachet services in 8th month for churned cutsomers

Summary of Logistics Regression (without Balancing)

Accuracy of LR without balancing on train data 0.9205714285714286

F1 Score of LR without balancing on train data 0.1070663811563169

Precision Score of LR without balancing on train data 0.6410256410256411

Recall Score of LR without balancing on train data 0.05841121495327103

	precision	recall	f1-score	support
0	0.92	1.00	0.96	8272
1	0.56	0.06	0.11	729
accuracy			0.92	9001
macro avg	0.74	0.53	0.53	9001
weighted avg	0.89	0.92	0.89	9001

Summary of Logistics Regression (with Balancing)

Accuracy of LR with weighted class balancing on train data 0.7074285714285714

F1 Score of LR with weighted class balancing on train data 0.3091972116033281

Precision Score of LR with weighted class balancing on train data 0.19145084934558618

Recall Score of LR with weighted class balancing on train data 0.8031542056074766

	precision	recall	f1-score	support
0	0.98	0.70	0.82	8272
1	0.19	0.81	0.31	729
accuracy			0.71	9001
macro avg	0.58	0.75	0.56	9001
weighted avg	0.91	0.71	0.78	9001

Summary of Logistics Regression (with SMOTE Balancing)

Accuracy of LR with SMOTE balancing on train data 0.7682237660721692

F1 Score of LR with SMOTE balancing on train data 0.7819215102807385

Precision Score of LR with SMOTE balancing on train data 0.73828934641426

Recall Score of LR with SMOTE balancing on train data 0.8310348403152219

	precision	recall	f1-score	support
0	0.98	0.71	0.82	8272
1	0.20	0.81	0.32	729
accuracy			0.72	9001
macro avg	0.59	0.76	0.57	9001
weighted avg	0.91	0.72	0.78	9001

Summary of Logistics Regression (with RFE & SMOTE Balancing)

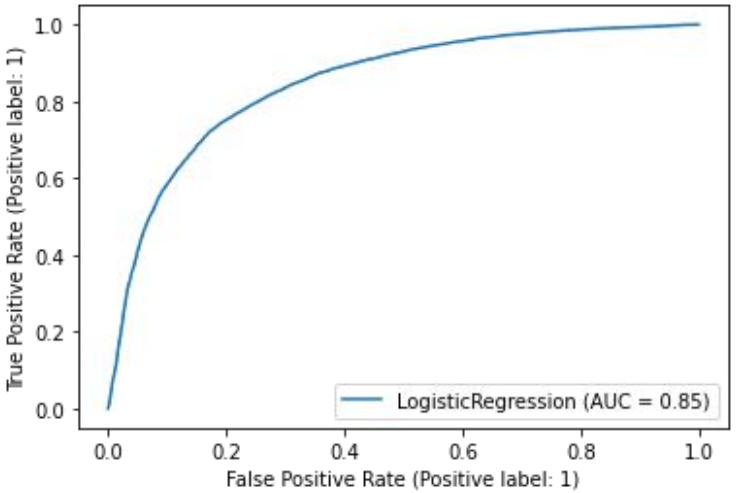
Accuracy of GLM with RFE and SMOTE balancing on train data 0.7666943177104936

F1 Score of GLM with RFE and SMOTE balancing on train data 0.780616224648986

Precision Score of GLM with RFE and SMOTE balancing on train data 0.7366580787633419

Recall Score of GLM with RFE and SMOTE balancing on train data 0.8301534632932394

	precision	recall	f1-score	support
0	0.98	0.70	0.82	8272
1	0.19	0.81	0.31	729
accuracy			0.71	9001
macro avg	0.58	0.75	0.56	9001
weighted avg	0.91	0.71	0.78	9001

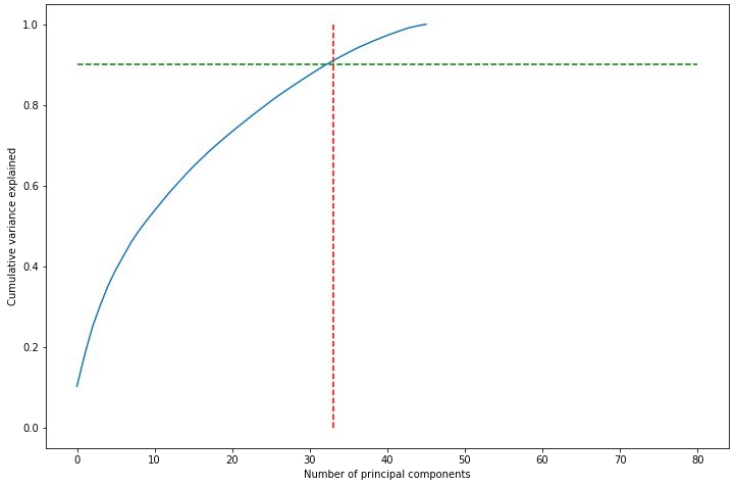


Summary of Logistics Regression (with Principal Component Analysis)

```
Accuracy of LR with PCA on train data 0.7684570717544588

F1 Score of LR with PCA on train data 0.7821888412017166

Precision Score of LR with PCA on train data 0.7383977900552486

Recall Score of PCA LR with on train data 0.8315014516798009
```

	precision	recall	f1-score	support
0	0.98	0.71	0.82	8272
1	0.20	0.81	0.32	729
accuracy			0.72	9001
macro avg	0.59	0.76	0.57	9001
weighted avg	0.91	0.72	0.78	9001

Optimum value of PCA component



Summary of Logistics Regression (with Optimum PCA)

Accuracy of LR with Otimum PCA on train data 0.7440377436748237

F1 Score of LR with Otimum PCA on train data 0.7586527180289402

Precision Score of LR with Otimum PCA on train data 0.7176748057713651

Recall Score of PCA LR with Otimum PCA on train data 0.8045935296557445

	precision	recall	f1-score	support
0	0.97	0.69	0.80	8272
1	0.18	0.76	0.29	729
accuracy			0.69	9001
macro avg	0.57	0.72	0.54	9001
weighted avg	0.91	0.69	0.76	9001

Summary of Decision Tree Classifier

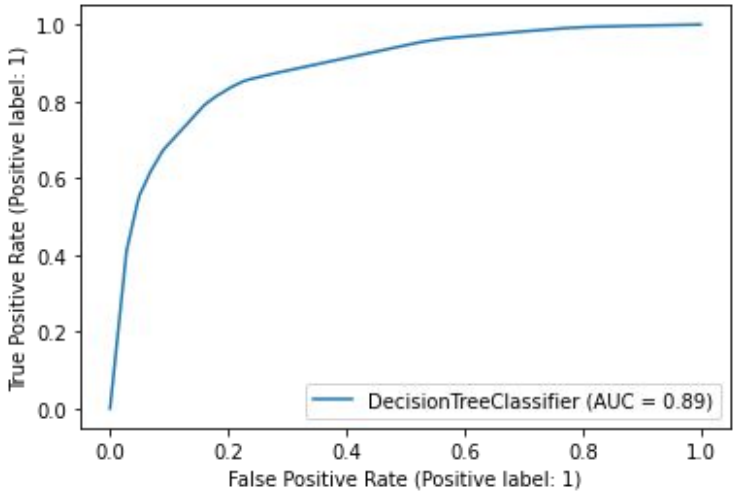
Accuracy of Decision Tree on train data 0.7684570717544588

F1 Score of Decision Tree on train data 0.7821888412017166

Precision Score of Decision Tree on train data 0.7383977900552486

Recall Score of Decision Tree on train data 0.8315014516798009

	precision	recall	f1-score	support
0	0.97	0.82	0.89	8272
1	0.25	0.70	0.37	729
accuracy			0.81	9001
macro avg	0.61	0.76	0.63	9001
weighted avg	0.91	0.81	0.85	9001



Score Summary of all models

	Model	Train Accuracy	Recall	Precision	F1_Score	Test Accuracy
0	LR w/o Balancing	0.9206	0.0584	0.6410	0.1071	0.9200
1	LR w.Class Balancing	0.7074	0.8032	0.1915	0.3092	0.7097
2	LR w.SMOTE Balancing	0.7682	0.8310	0.7383	0.7819	0.7160
3	GLM/RFE	0.7667	0.8302	0.7367	0.7806	0.7100
4	PCA	0.7685	0.8315	0.7384	0.7822	0.7159
5	Optimum PCA	0.7440	0.8046	0.7177	0.7587	0.6915
6	DT	0.7685	0.8315	0.7384	0.7822	0.8085
7	Tuned DT	0.7685	0.8315	0.7384	0.7822	0.8474
8	RF	1.0000	1.0000	1.0000	1.0000	0.9136
9	Tuned RF	0.9793	0.9868	0.9723	0.9795	0.9058
10	XGB	0.8939	0.8913	0.8959	0.8936	0.8739
11	Tuned XGB	0.9403	0.9370	0.9432	0.9401	0.8739

Conclusion

- Decrease in Total data recharge amount and Maximum recharge amount in month 8 is a strong indicator of churn.
- Decrease in 2G usage for Month 8 shows an increase trend of churn.
- Decrease in incoming and outgoing special calls in month 8 shows high churn probability.
- In general overall decrease in all kind of outgoing calls indicates a potential churn.
- New customers tend to churn easily.
- XGBoost and Random Forest produced the best prediction scores.
- Logistic Regression with SMOTE balancing produced the best interpretable model with 29 variables.