TELECOM CHURN CASE STUDY

GROUP MEMBERS:

- ***Mr. Abhinav***
- ***Mr. Rohit Mathur***
- ***Mr. Neha Yadav***

PROBLEM STATEMENT

In the telecom industry, customers can 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. Hence, it is vital for us to understand which customers may churn so that the rentention strategies can be devised accordingly.

BUSINESS OBJECTIVE

In this project, our business objective is to analyse customer-level data of a leading telecom firm, build predictive models to identify customers at high risk of churn and identify the main indicators of churn. Thus, our focus would be on

- **Retaining high profitable customers.**
- **Predicting which customers are at high risk of churn (in order to devise customer retention strategies accordingly).**

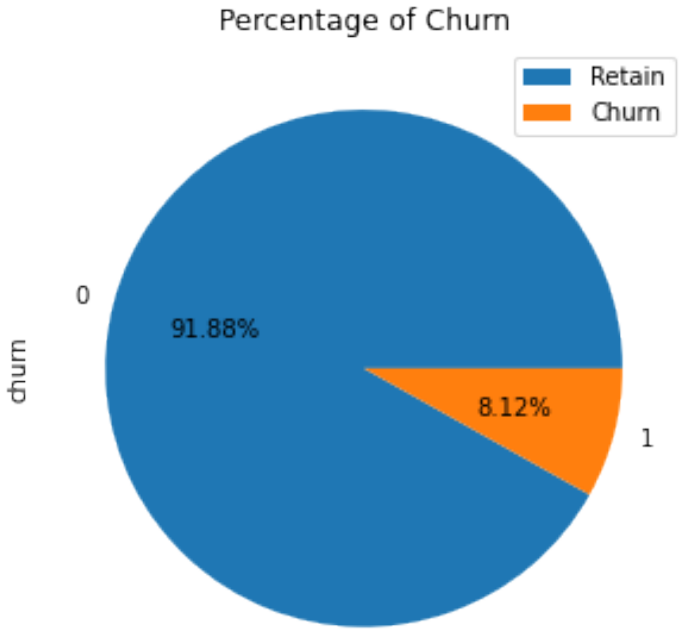
SOLUTION METHODOLOGY

- Data Understanding, cleaning and data manipulation.
  - Check and handle duplicate data.
  - Check and handle NAN values and missing values.
  - Drop columns, if it contains large number of missing values and not useful for the analysis.
  - Imputation of the values, if necessary.
  - Check and handle outliers in data.
- EDA
  - Univariate data analysis: value count, distribution of variable etc.
  - Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Feature scaling & dummy variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

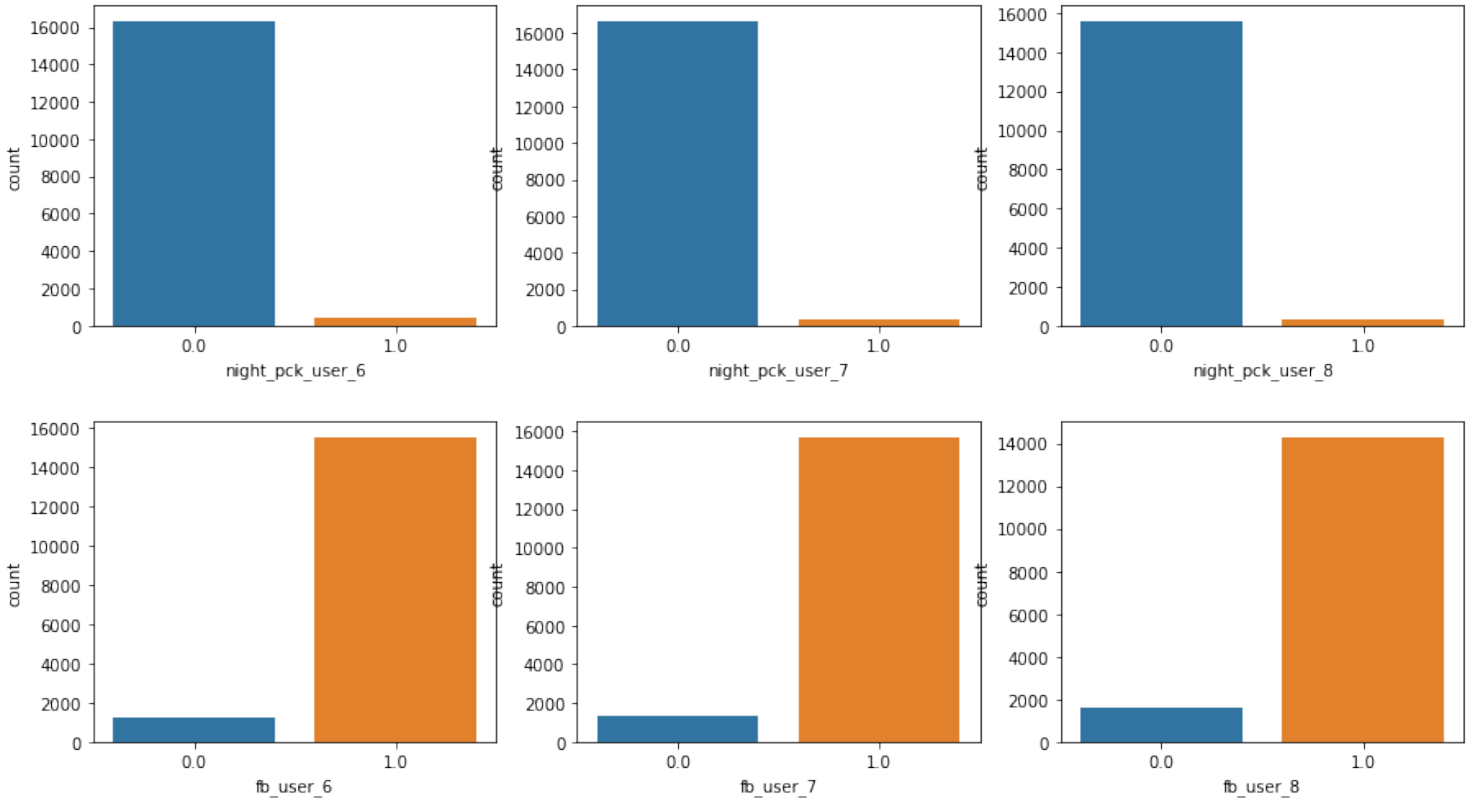
DATA MANIPULATION

- Total Number of Rows =99,999, Total Number of Columns =226.
- In churn prediction, we are given that there are three phases of customer lifecycle :
  - (i) The ‘good’ phase [Month 6 & 7]: First two months
  - (ii) The ‘action’ phase [Month 8]: Third month
  - (iii) The ‘churn’ phase [Month 9]: Fourth Month
- Among the 'rech_amt' and 'rech_data' features, all the data features have around 74% missing values whereas the calling based recharge features have no missing values.
- For columns like av_rech_amt_data_* and total_rech_data_* (months i.e 6,7,8&9) the minimum value is either 1 or 0.5. Hence, we can impute the missing values by 0. (Considering there were no recharges done by the customer).
- On filtering the high-value customers, we get about 29.9K observations. These records of high-value customers shall be used for further analysis.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- There are no missing values in total_og_mou_9, total_ic_mou_9, vol_2g_mb_9, vol_3g_mb_9 features of churn phase.

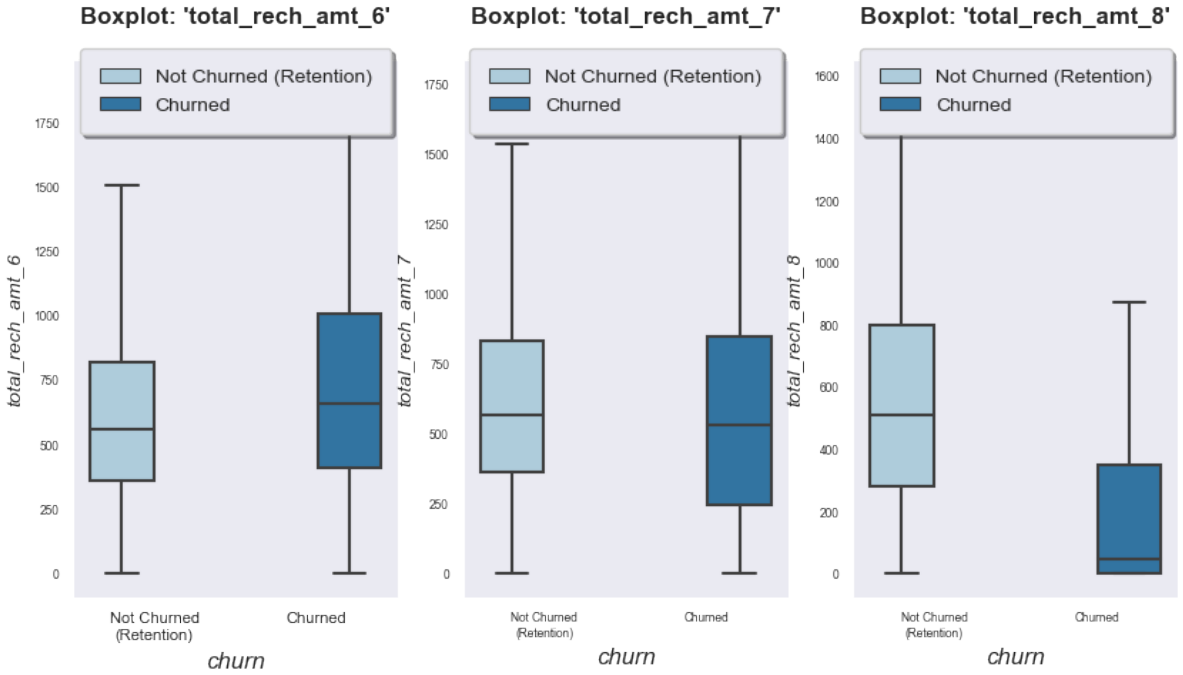
EDA



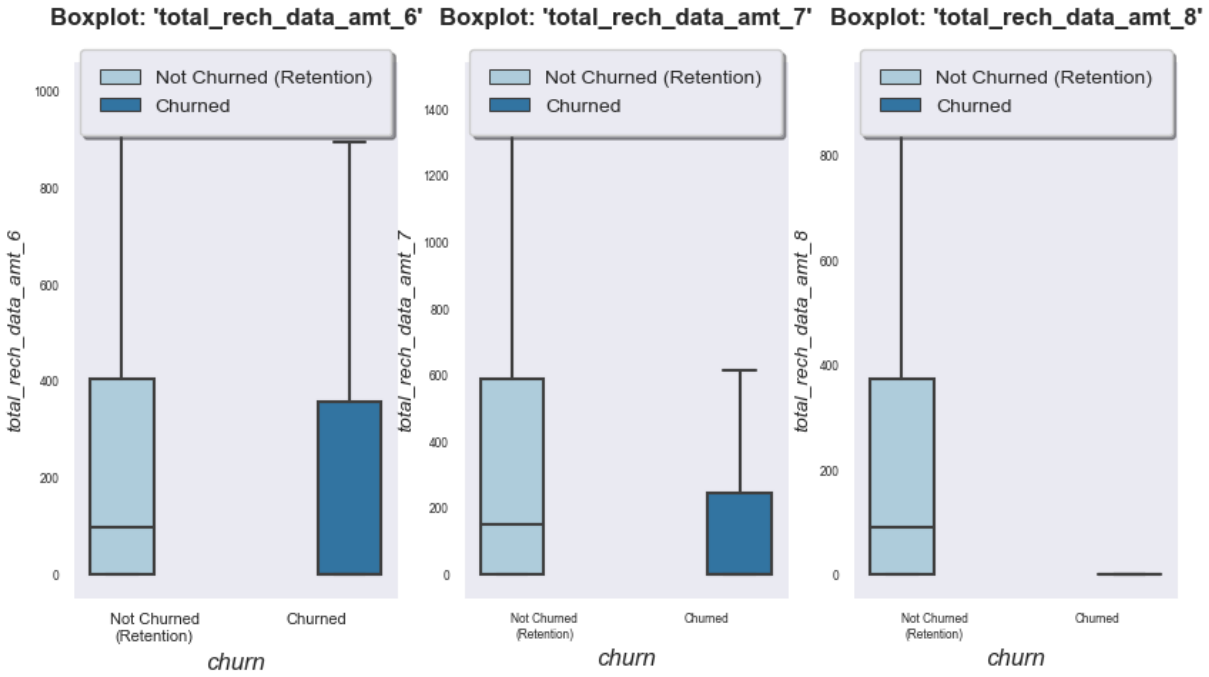
Checking the % of churn in the entire dataset



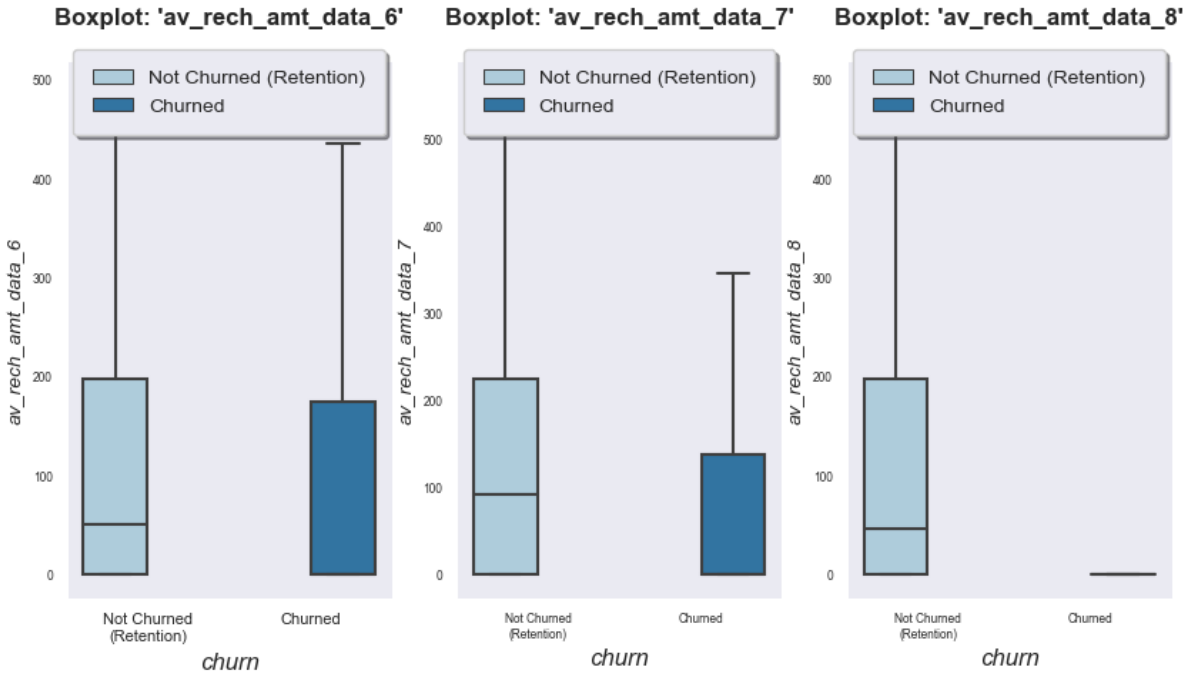
Univariate Analysis on categorical columns



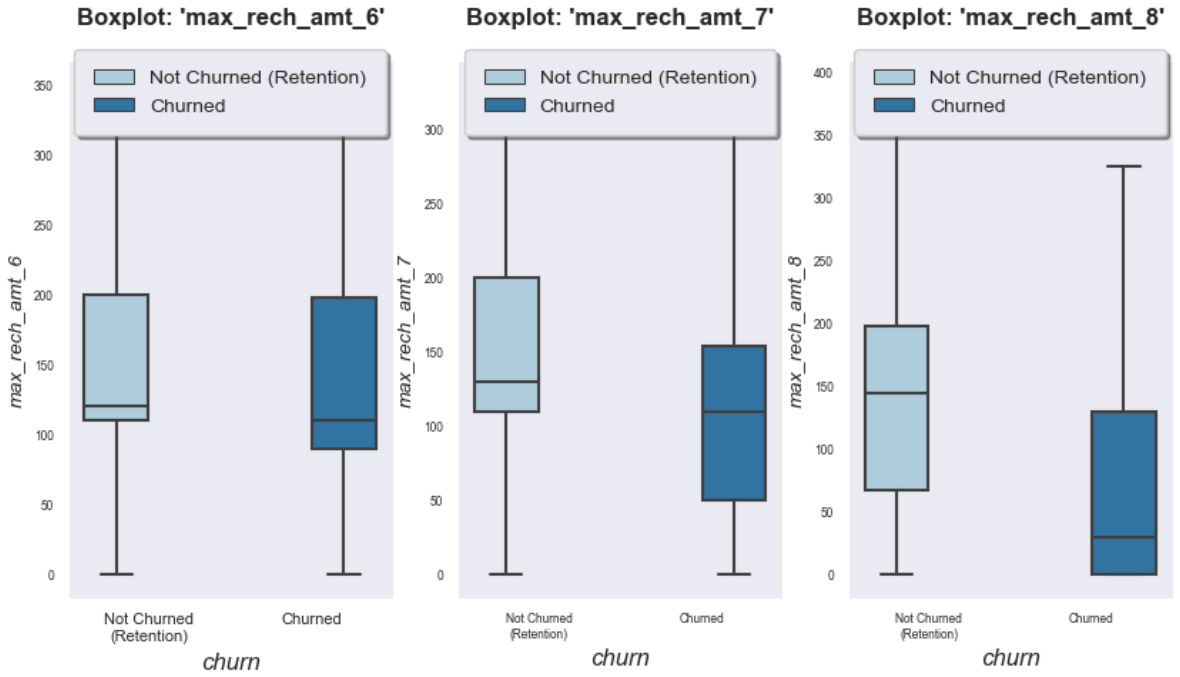
Observation: We can see a drop in the total recharge amount for churned customers in the 8th Month (Action Phase).



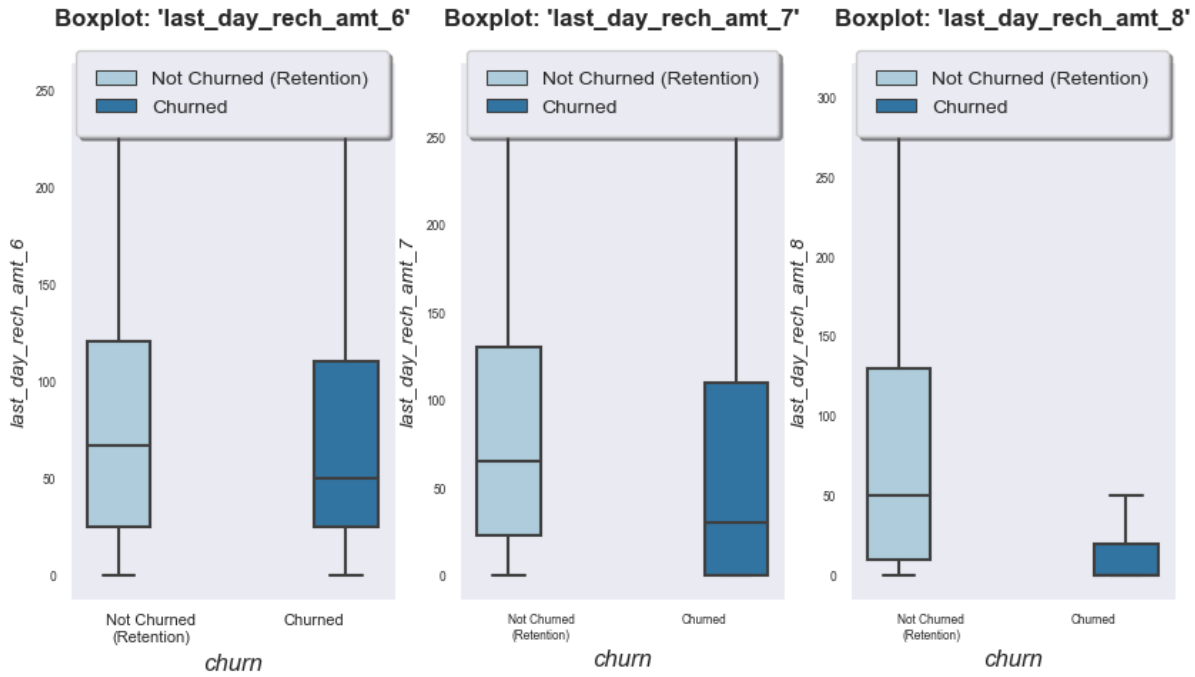
Observation: Again, a significant drop in the total data recharge amount is seen for churned customers in the 8th Month (Action Phase).



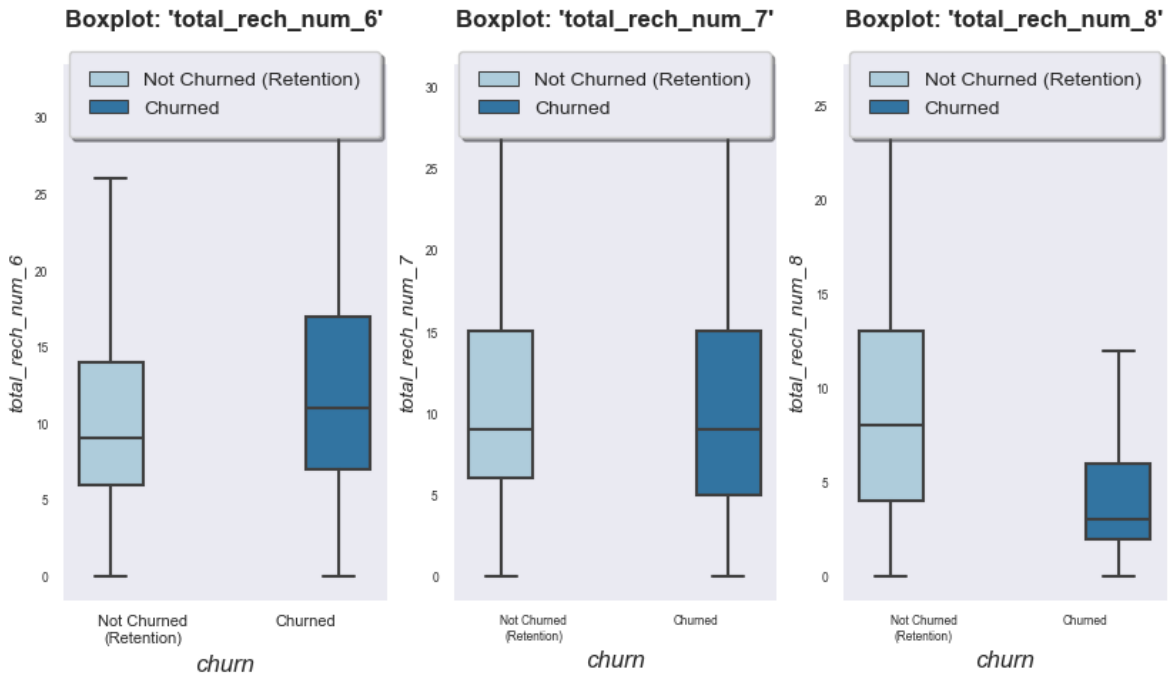
Observation: Again, a significant drop in the average data recharge amount is seen for churned customers in the 8th Month (Action Phase). Also, it makes sense as the total_rech_data_amt_* drops too.



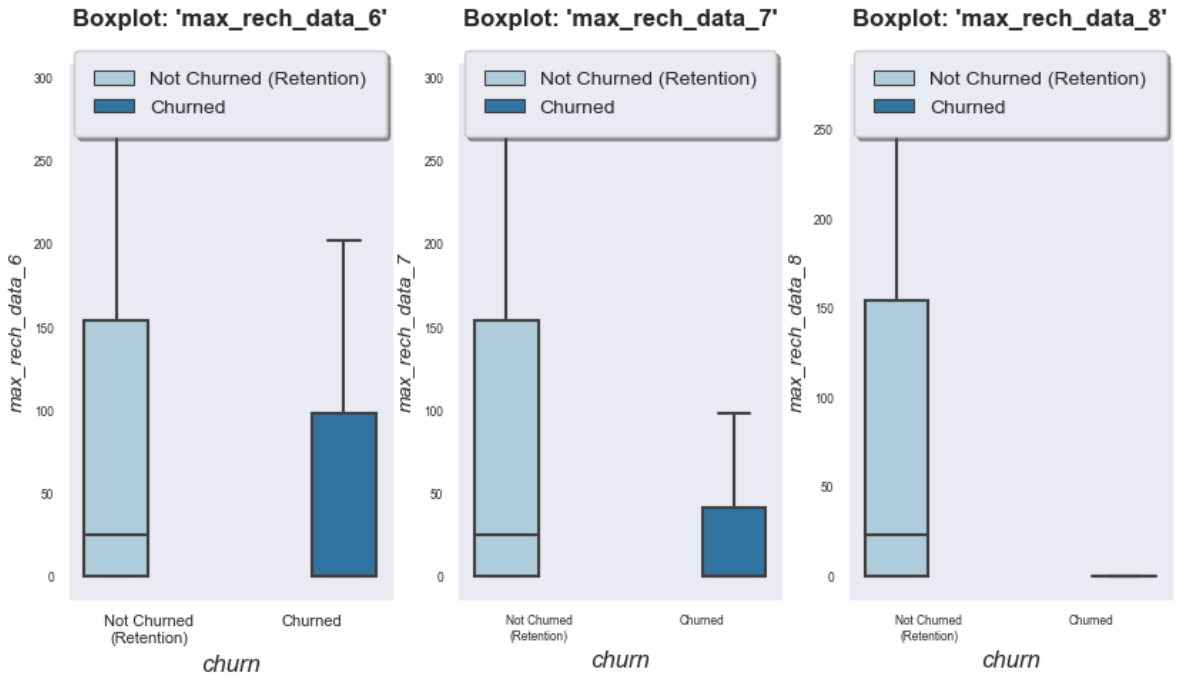
Observation: A noticeable drop in the maximum recharge amount is seen in the 8th month (action phase) for churned customers.



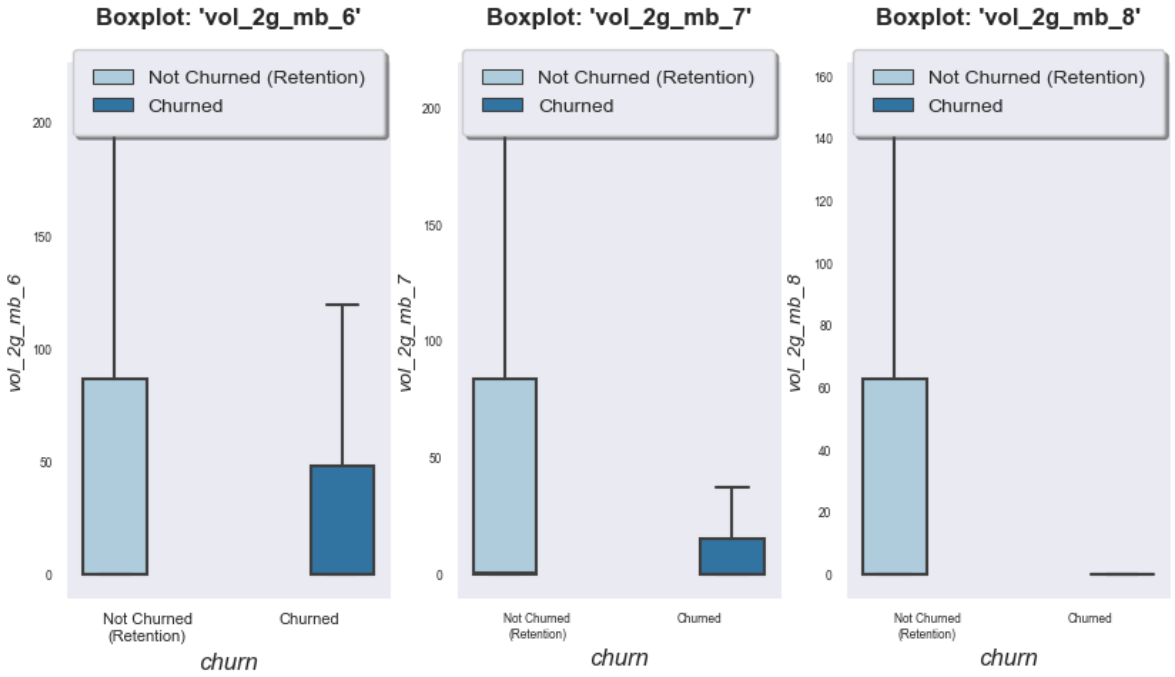
Observation: We do see a noteworthy drop in the last day recharge amount in the 8th month (action phase) for churned customers.

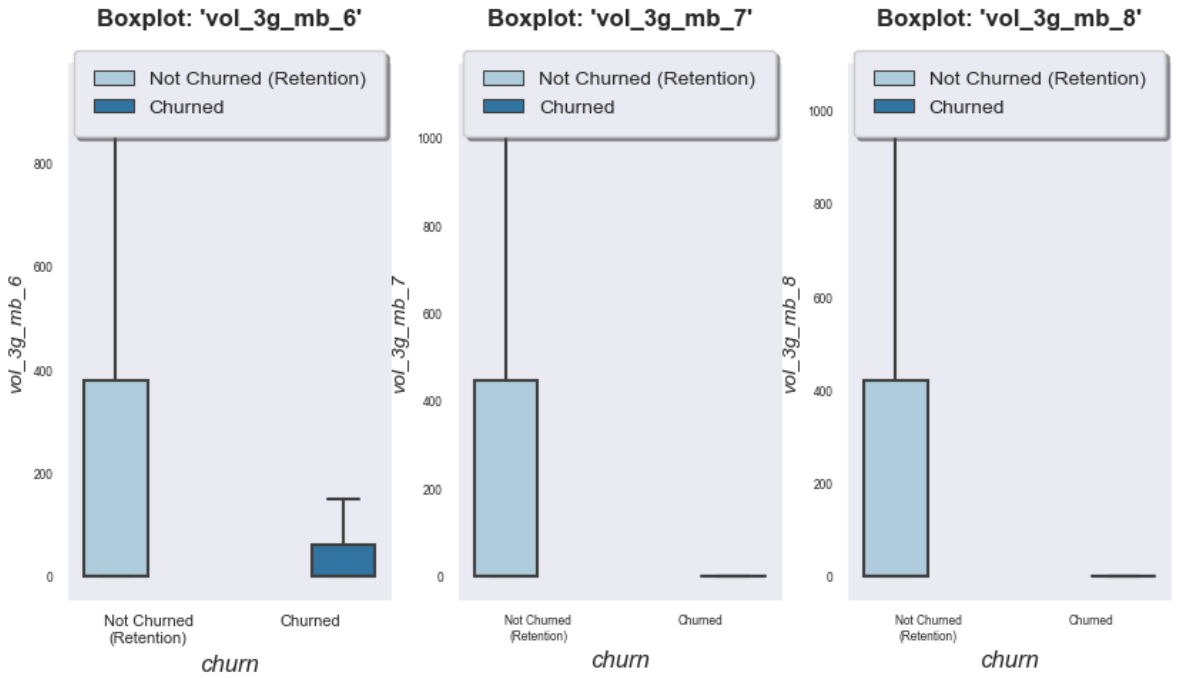


Observation: A noticeable drop in the total number of recharges is seen in the 8th month (action phase) for churned customers.



Observation: A significant drop in the max_rech_data is seen in the 8th month (action phase) for churned customers.

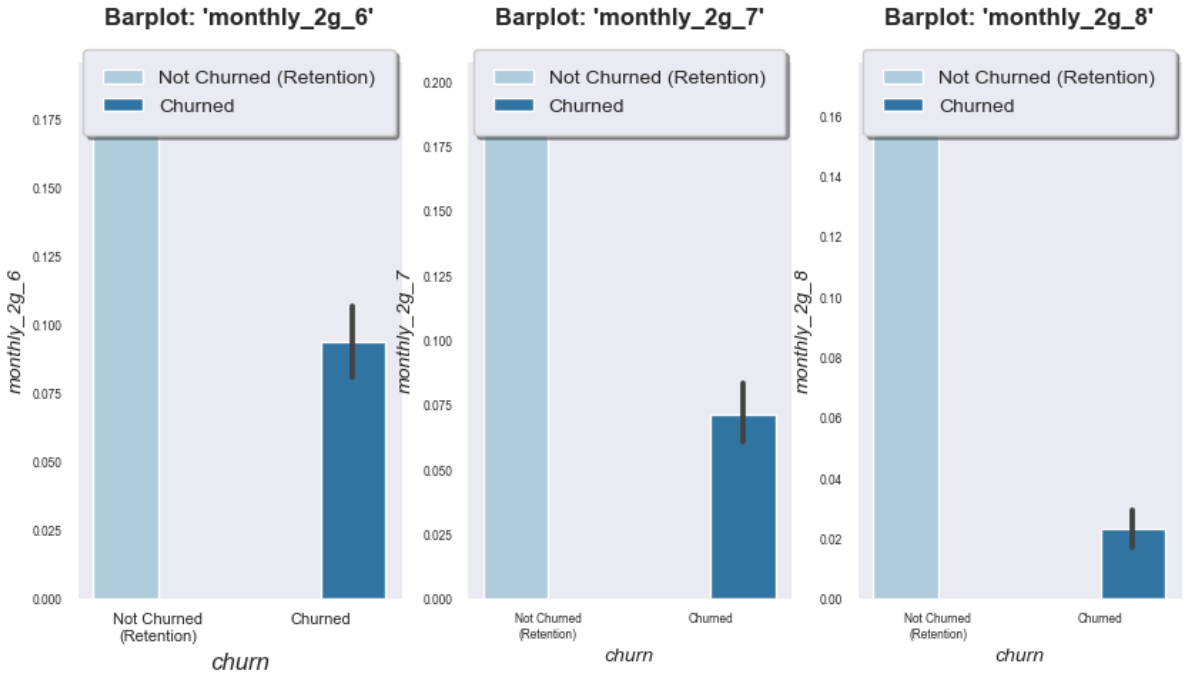


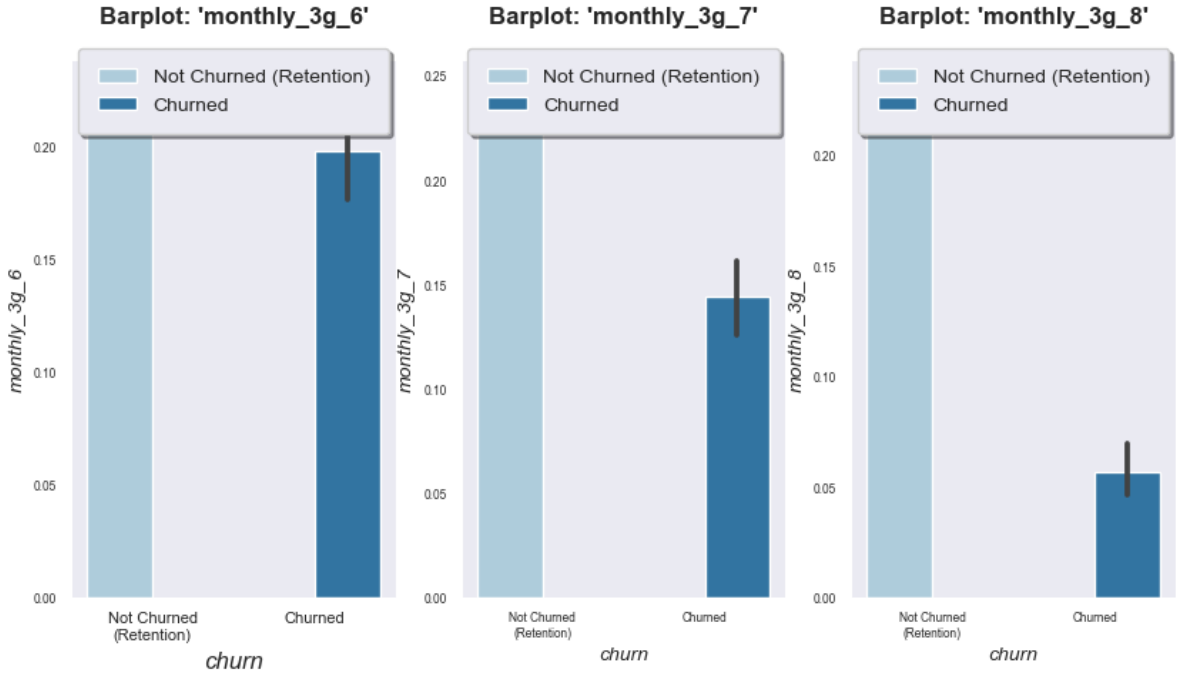


Observations:

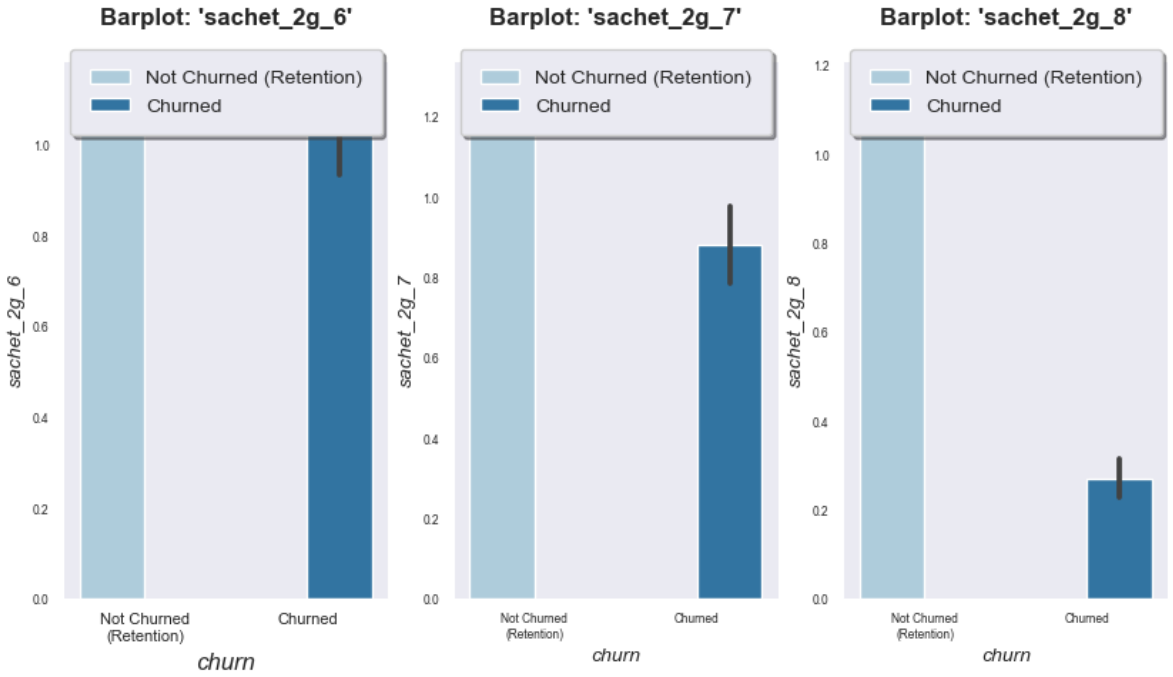
(i) The volume of 2G and 3G data usage substantially drops in the 8th month(action phase) for churned customers.

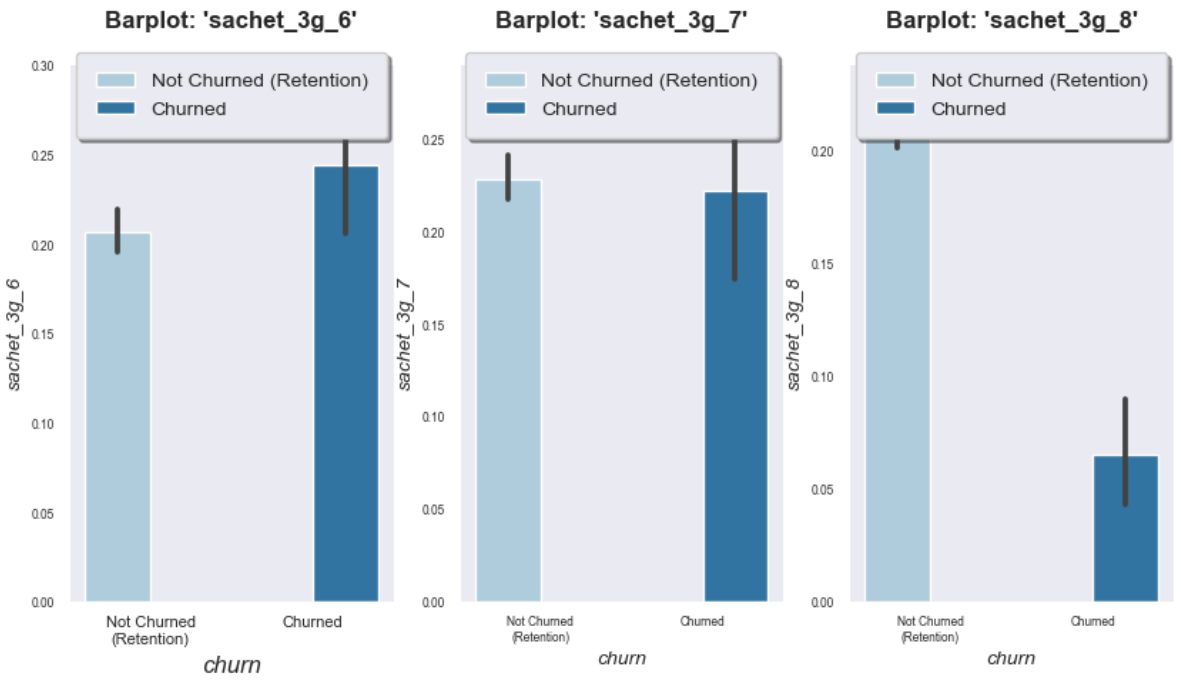
(ii) Also, we see the usage of 2G data is comparatively lesser than that of 3G data, though the drop seems to follow similar pattern.



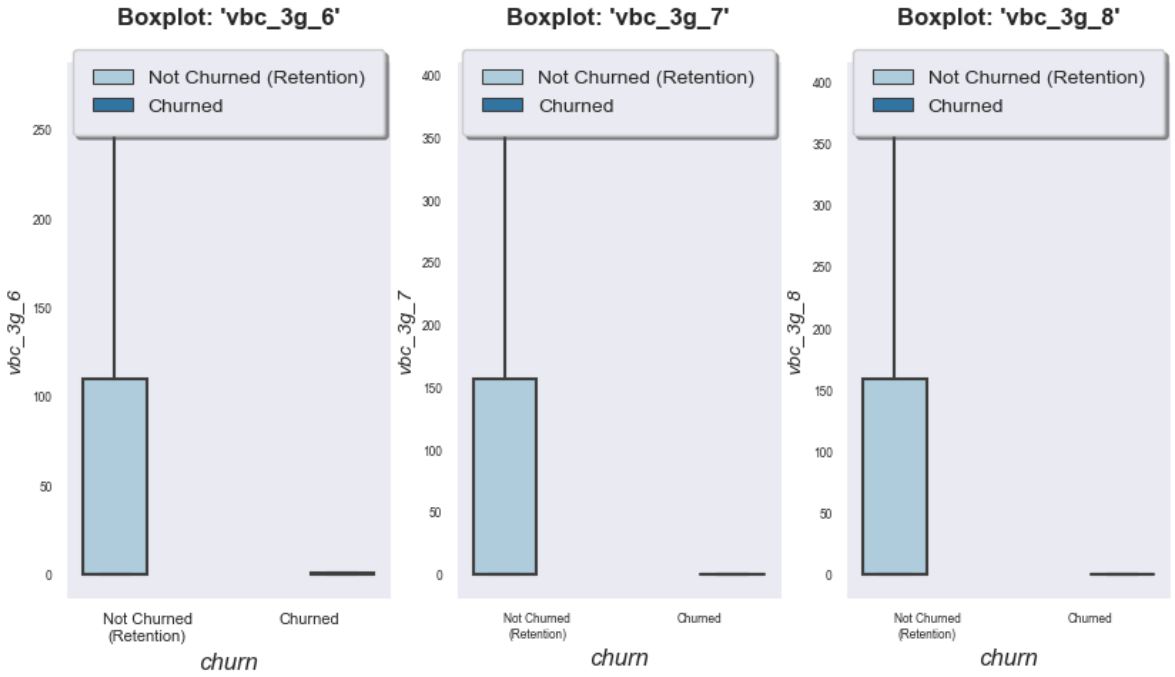


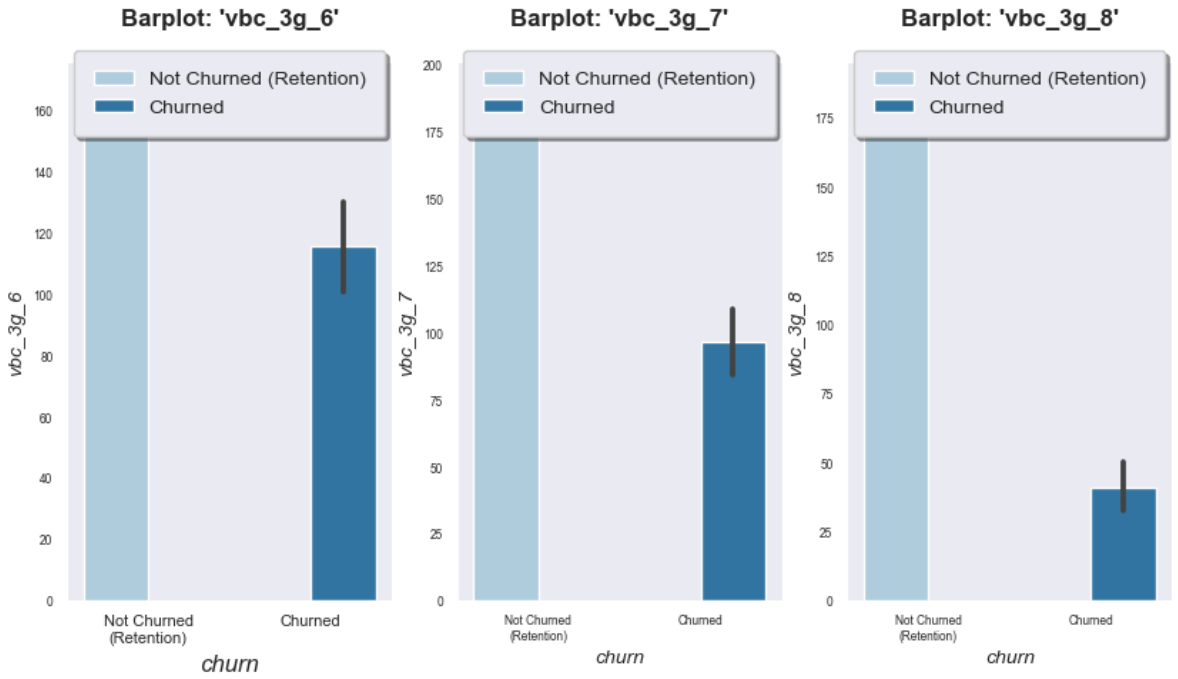
Observation: Again, we can see a drop in monthly 2G and 3G subscriptions for churned customers in 8th Month (action phase).





Observations: We see that the 'sachet_2g' and 'sachet_3g' schemes we largely used in the first and second months of good phase i.e. month (6) and (7) and then the trend shows a sudden drop in the usage as we approach the 8th month (action phase) for churned customers.

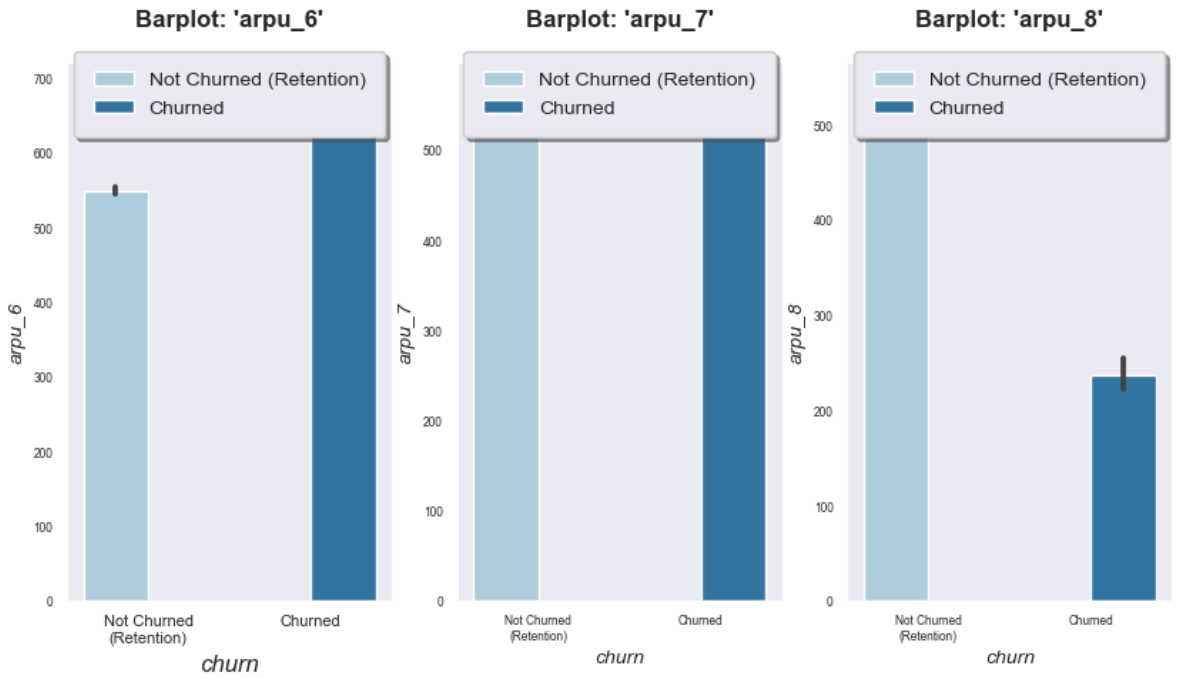




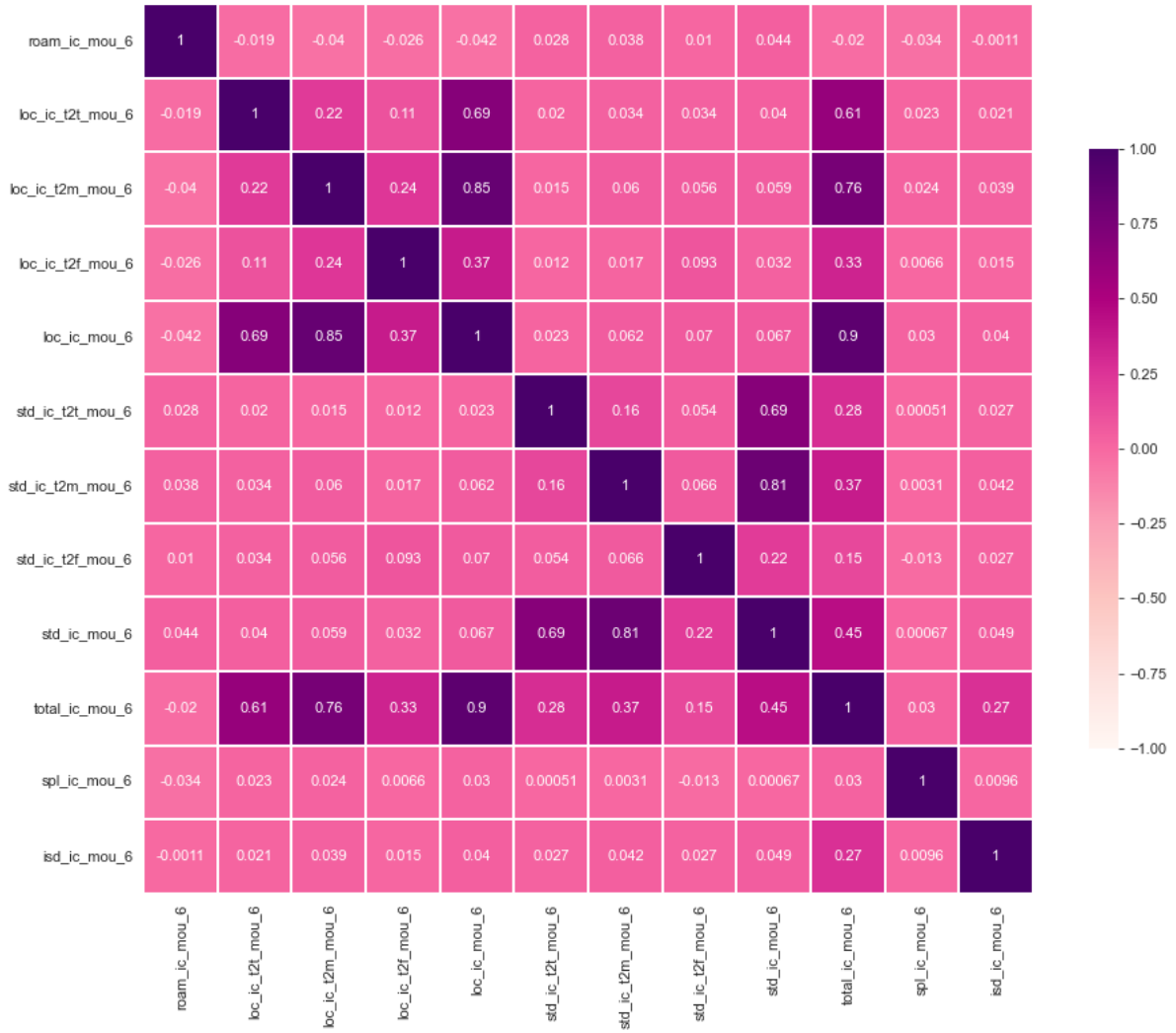
Observations:

(i) The volume-based cost, vbc for 3G is much lower for Churned customers than the non-churned customers.

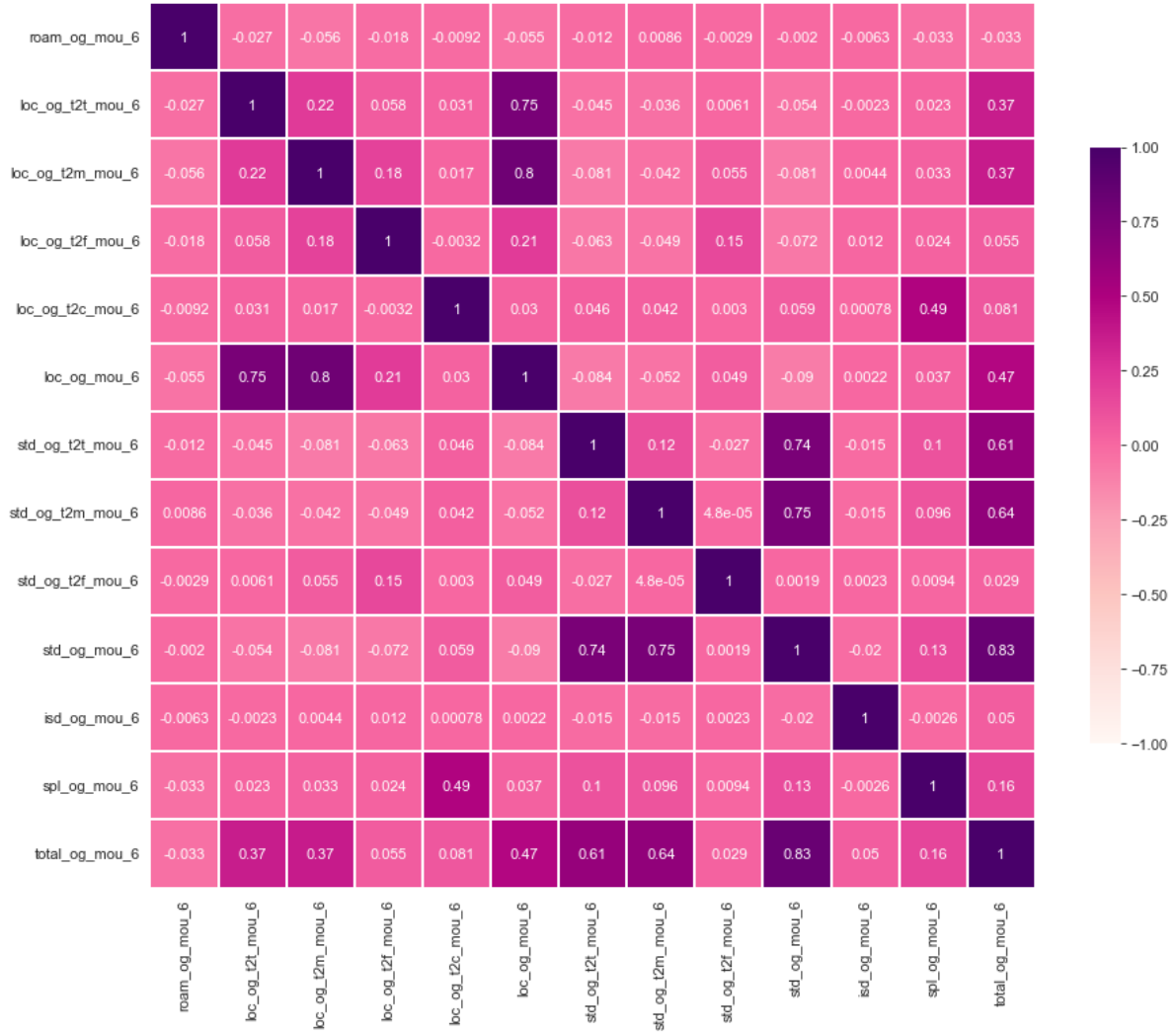
(ii) Also, we see a drop in the vbc as we approach the 8th month(action phase) for churned customers.



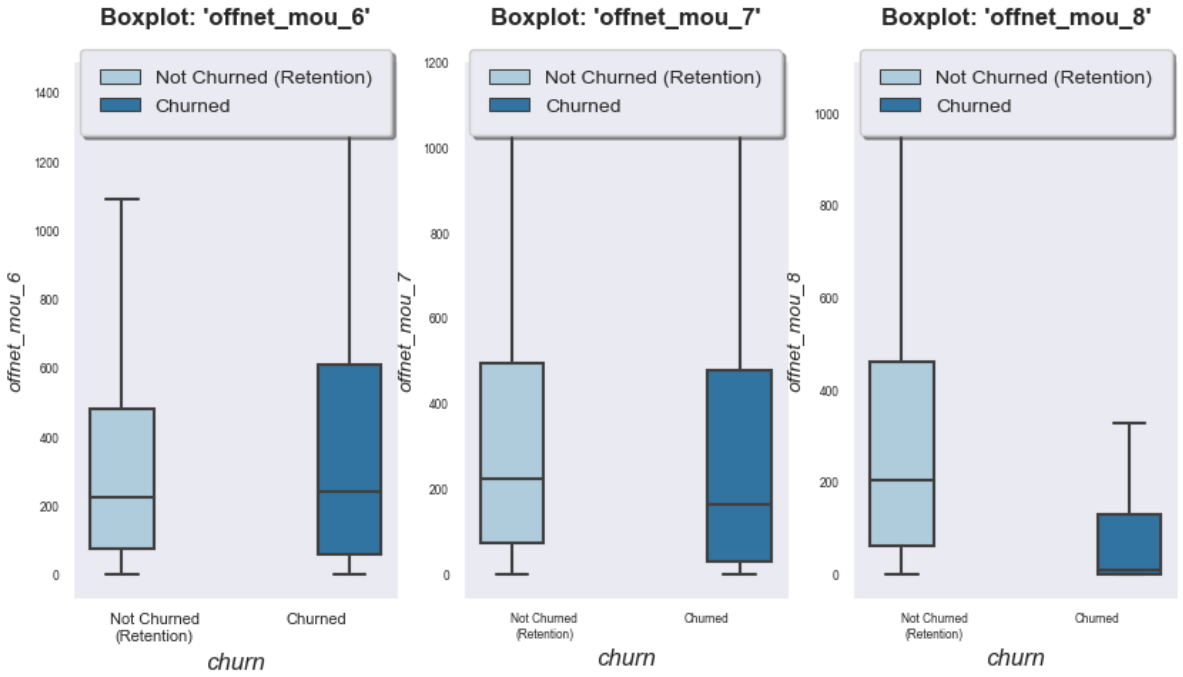
Observation: We see a drop in the average revenue per user for churned customers as we approach month 8(action phase) from the good phase (i.e. months 6 and 7)

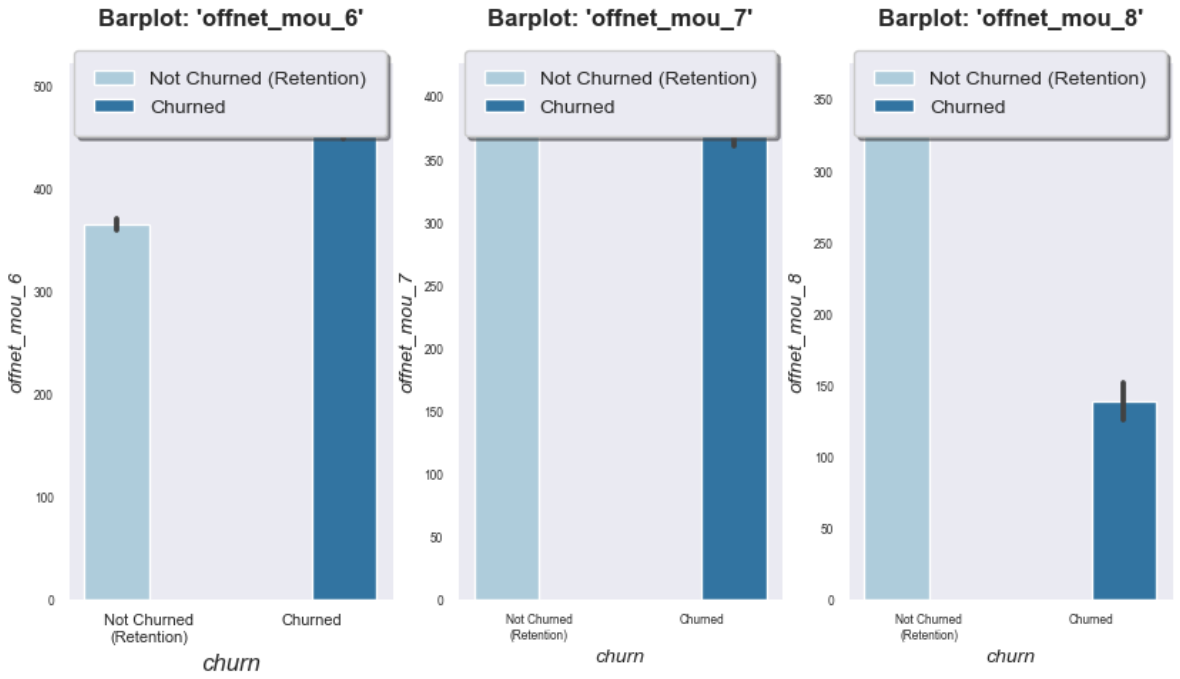


- In above correlation matrix, total_ic_mou_6, std_ic_mou_6 and loc_ic_mou_6 seems to have strong correlation with other fields, and they needs to be inspected to avoid any multicollinearity issues.

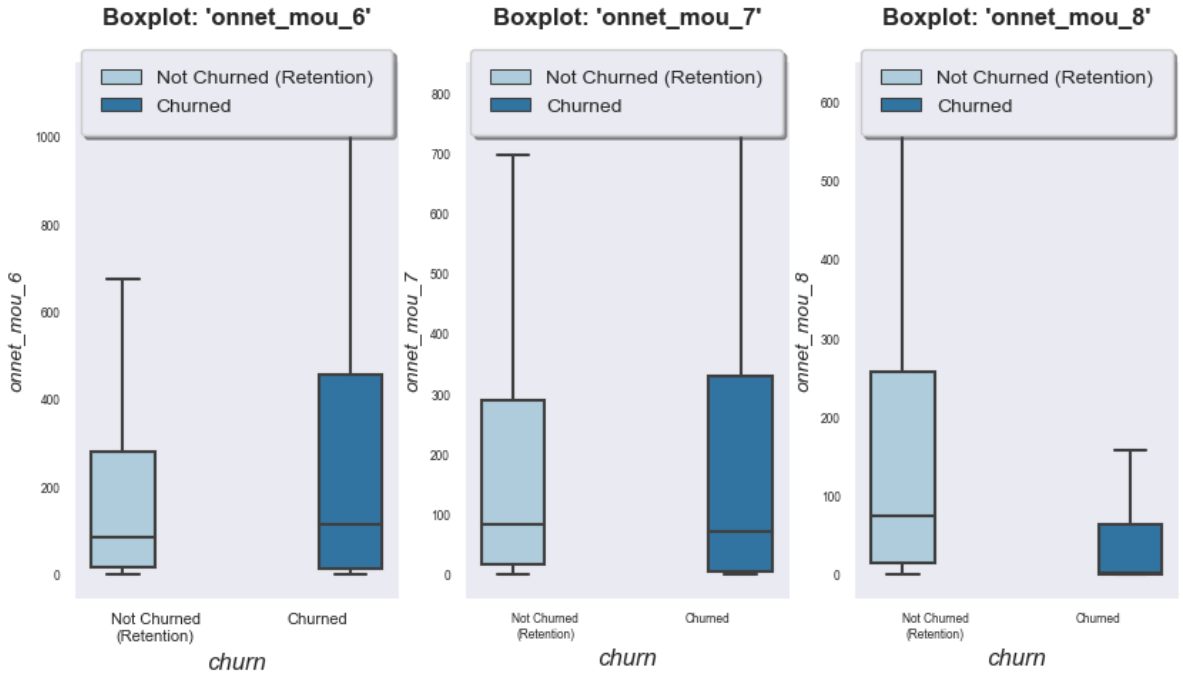


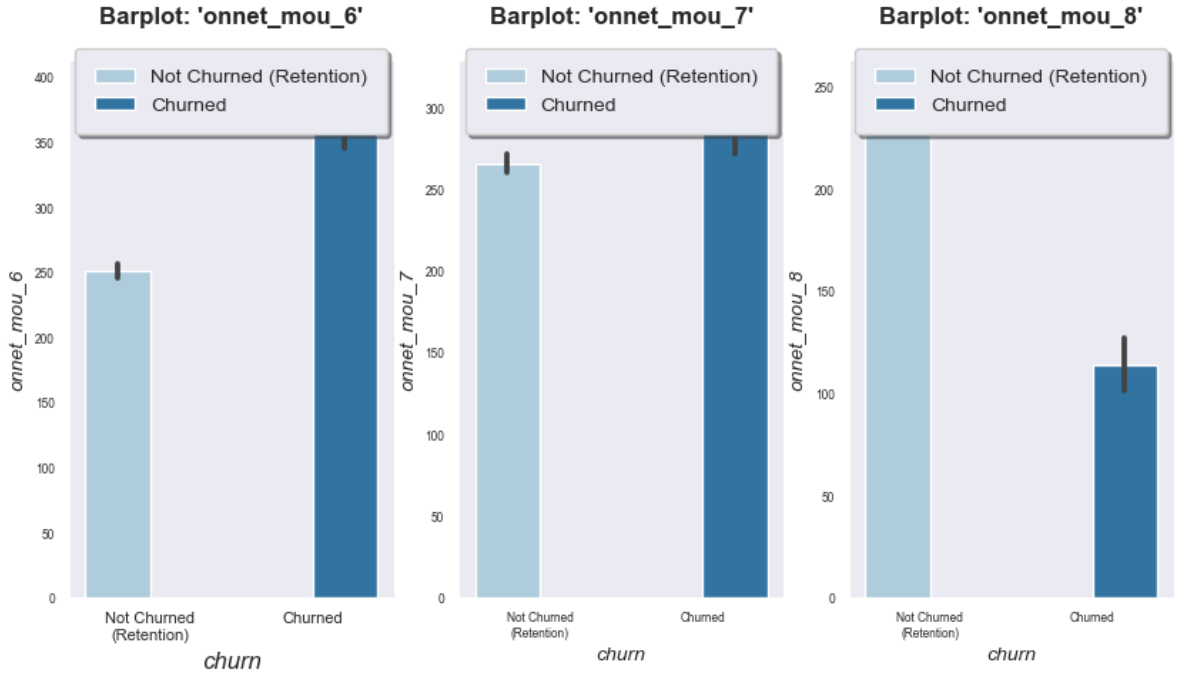
- In above correlation matrix, total_og_mou_6, std_og_mou_6 and loc_og_mou_6 seems to have strong correlation with other fields, and they needs to be inspected to avoid any multicollinearity issues



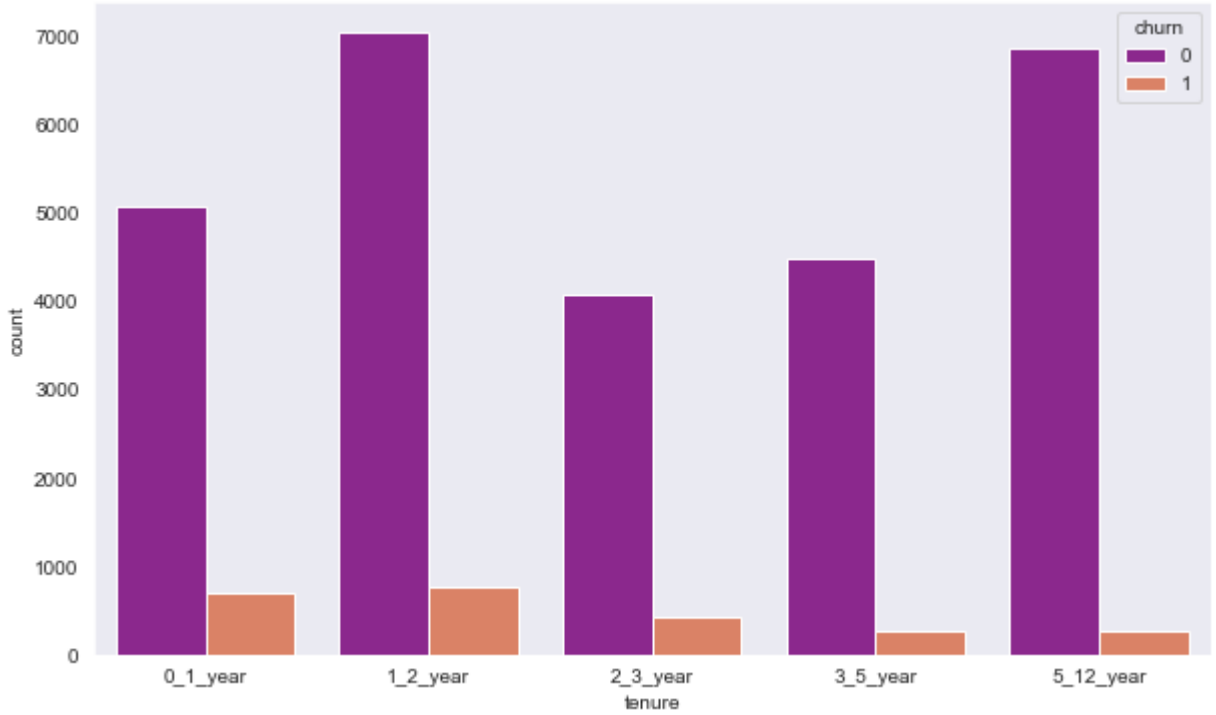


Observation: The offnet_mou (minutes of usage) decline as we move from good phase (i.e. month 6 and 7) to the action phase(month 8) for churned customers.





Observation: The onnet_mou (minutes of usage) decline as we move from good phase (i.e. month 6 and 7) to the action phase(month 8) for churned customers.

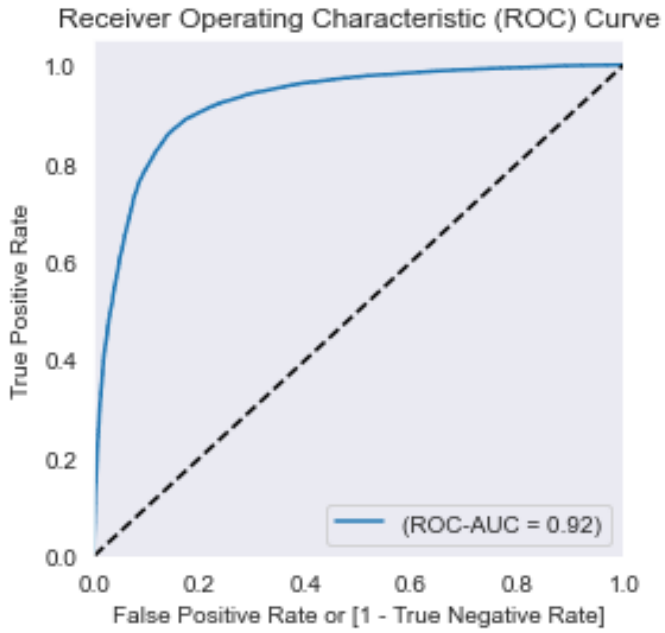


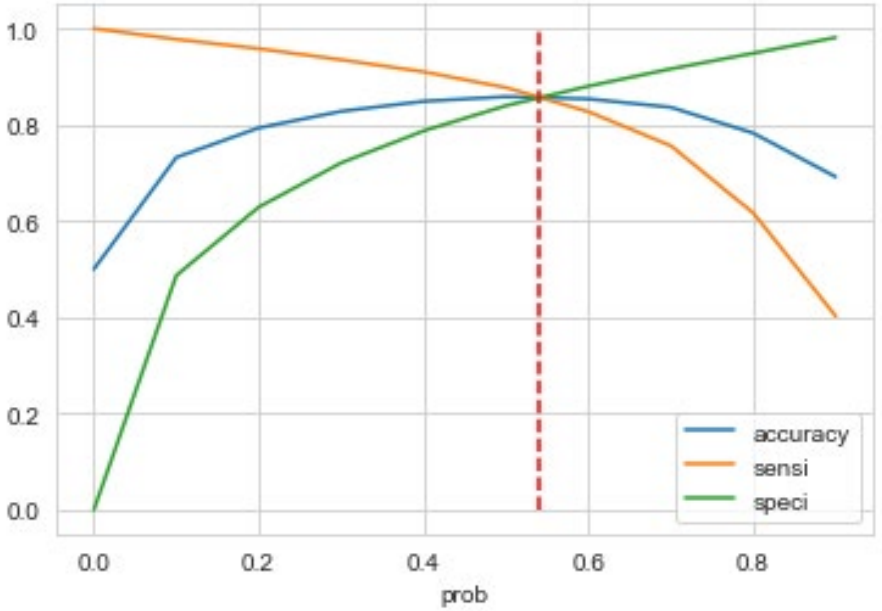
Observation: We see that as the age on network (AON) increases, the number of people churned out (i.e. churned class label: 1) are reduced.

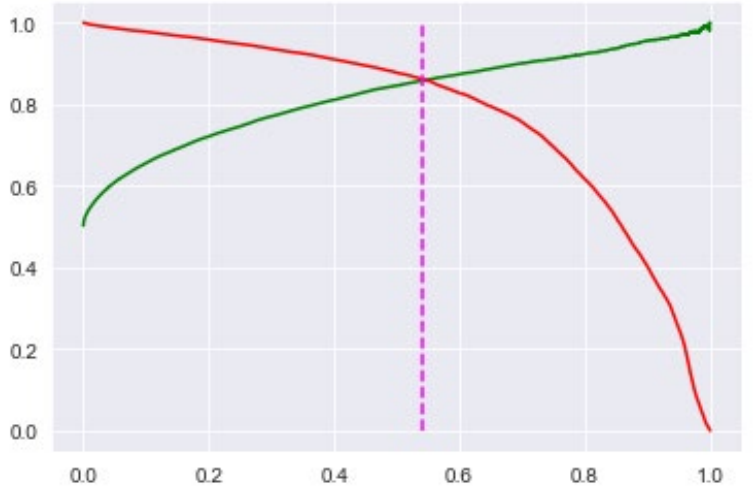
MODEL BUILDING

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 51 variables as output
- Building Model by removing the variable whose p-value is greater than 0.05 and vif value is greater than 5
- Predictions on test data set
- Overall accuracy 86%

ROC CURVE-1



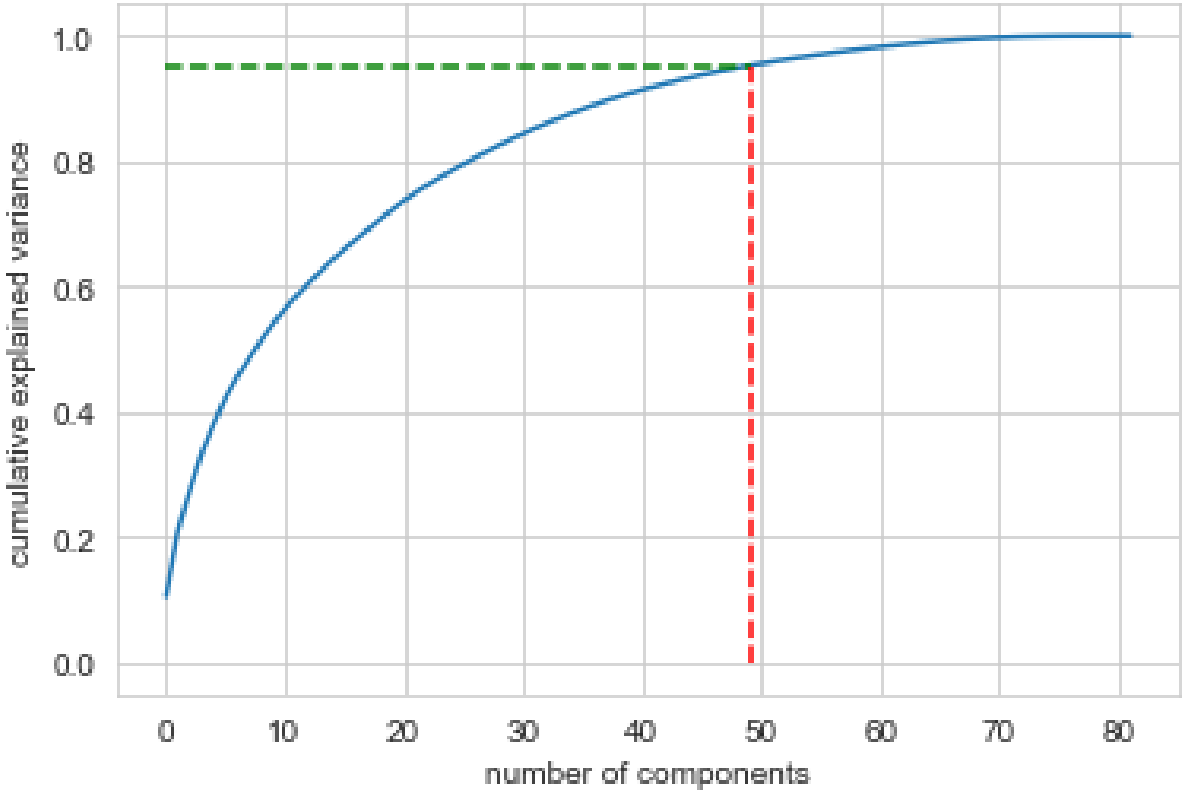




Precision and Recall Trade-off

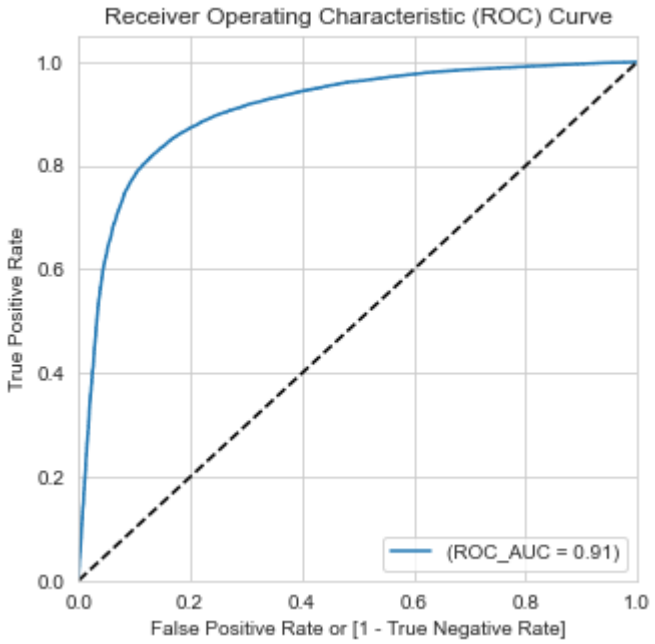
Finding Optimal Cut off Point

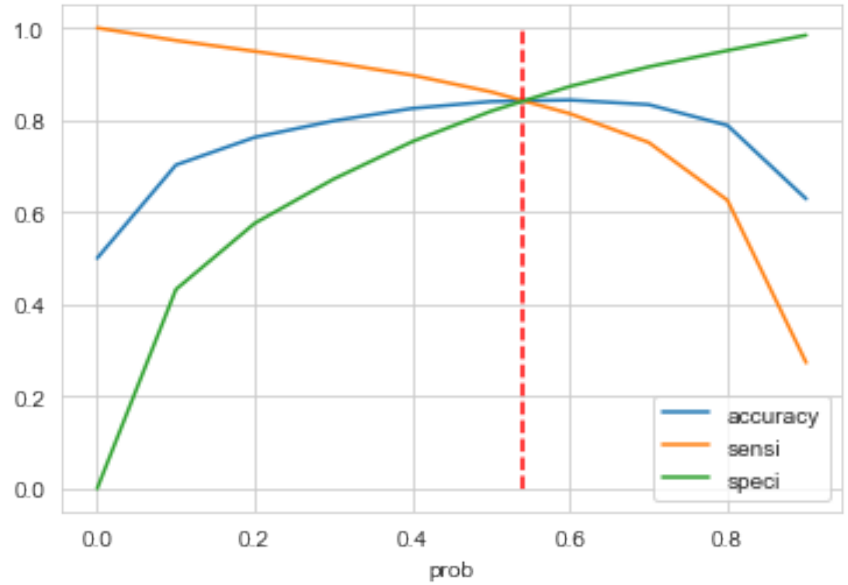
- **Optimal cut off probability is that probability where we get balanced sensitivity and specificity.**
- **From the second graph it is visible that the optimal cut off is at 0.54.**
- **ROC_AUC of the train set: 0.923**
- **Recall Score of the train set: 0.86**

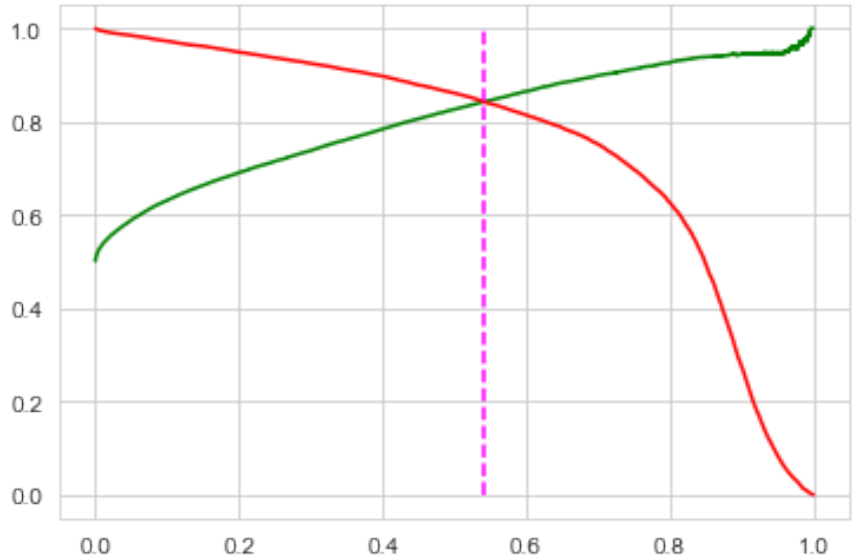


Cumulative Variance Vs Number of Components

ROC CURVE-2





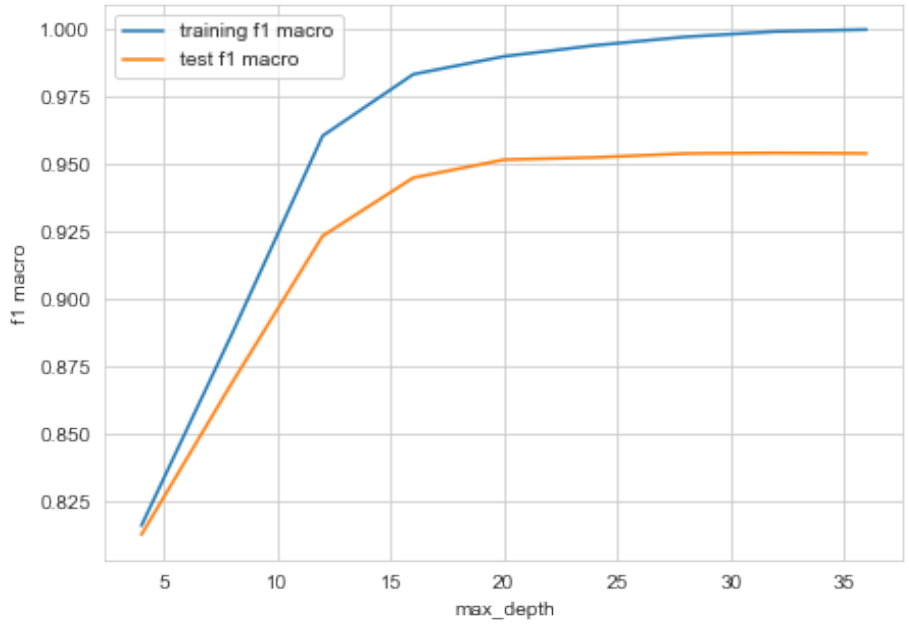


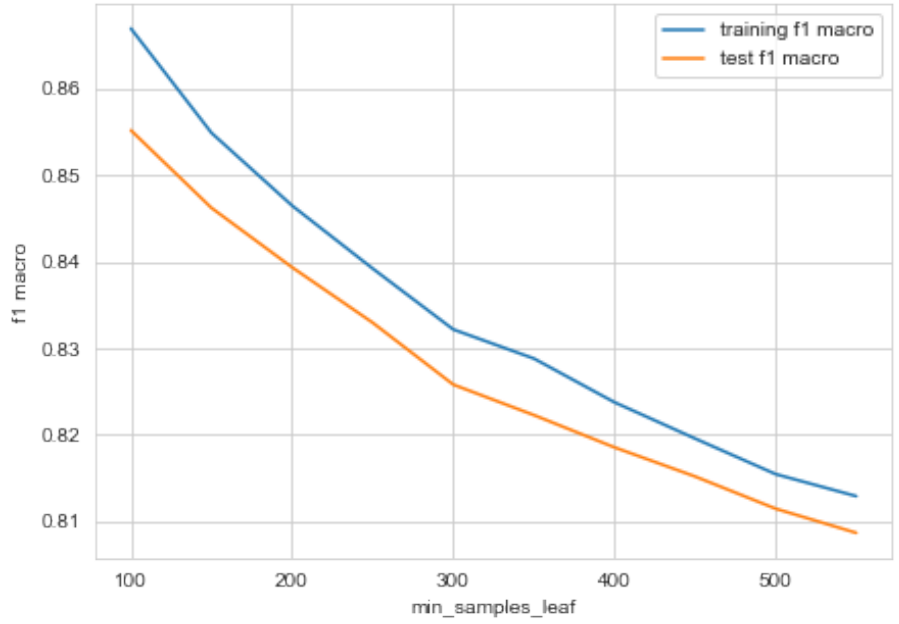
Precision and Recall Trade-off

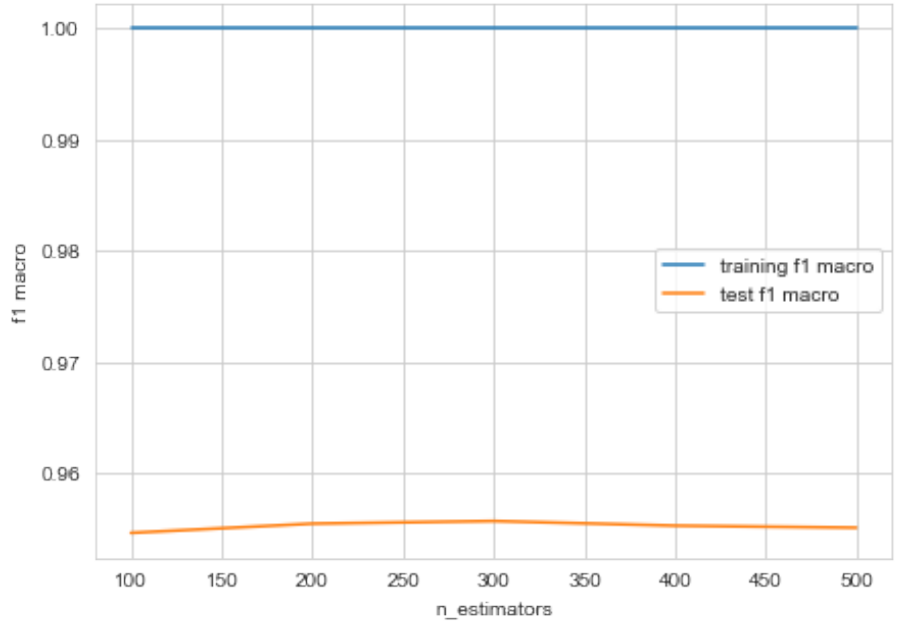
Finding Optimal Cut off Point

- **Optimal cut off probability is that probability where we get balanced sensitivity and specificity.**
- **From the second graph it is visible that the optimal cut off is at 0.54.**
- **ROC_AUC of the train set: 0.91**
- **Recall Score of the train set: 0.84**

F-1 Score of Train & Test Data







FEATURE ANALYSIS:

- **Age on the network (tenure) is an important factor that determines the churn in High value customers; the longer a customer stays on the network lesser are the chances of him/her churning out.**

- **Total number of recharges done in the action month(8th month) is an important indicator for churn of high value customers where less number of recharges in action month are a good indicator for customer churning.**

- **Minutes of usage (Voice) in the action month combined is a very good indicator for customer churn. As the MOU in 8th month decreases, the chances of the customer churning out increases.**

- **Churn customers seems to have a relatively High ARPU (Average Revenue Per User) in the good phase, indicating that a sudden downturn in the ARPU from good to action phase is a major indicator of the customer churn.**

- **Total and Max data recharges for customers that churn out seem to be on a lower side as compared to that of non churned customers.**

FEATURE ANALYSIS:

Rank (Feature Importance Based)	Features	Coefficients
1.0	tenure_3_5_year	-2.1416
2.0	tenure_5_12_year	-2.1412
3.0	tenure_2_3_year	-1.7076
4.0	tenure_1_2_year	-1.4474
5.0	loc_ic_t2m_mou_8	-0.9354
6.0	arpu_avg_6_7	0.7385
7.0	spl_ic_mou_8	-0.6880
8.0	total_rech_num_8	-0.6425
9.0	total_rech_data_8	-0.5450
10.0	last_day_rech_amt_8	-0.5336
11.0	std_ic_t2t_mou_8	-0.5207
12.0	loc_ic_t2t_mou_8	-0.4978
13.0	loc_og_t2m_mou_8	-0.4842
14.0	max_rech_data_8	-0.3918
15.0	loc_ic_t2f_mou_8	-0.3857

BUSINESS RECOMMENDATIONS:

- **Recently joined customers (Age on Network: 0-2 years) can be provided addon incentives for a fixed period of time.**
- **Provide recharge incentives(data + voice) to high value customers in the action phase to help drive customer retention.**
- **Customer with high ARPU in good phase can be provided usage-based incentives to drive up the ARPU in action phase.**
- **Can provide free or discounted local onnet and mobile usage voice minutes during the action phase.**

PREDICTIONS SUMMARY:

Model No.	Models	ROC_AUC (Train)	Recall (Train)	ROC_AUC (Test)	Recall (Test)
1	Logistic Regression (RFE)	0.9230	0.86	0.8850	0.76
1.1(i)	Logistic Regression (PCA)	0.9100	0.84	0.8960	0.80
1.1(ii)	Logistic Regression (PCA+Hyperparameter Tuning)	0.9097	0.84	0.8962	0.80
1.2	Random Forest (PCA+Hyperparameter Tuning)	0.9110	0.84	0.8790	0.77

- **BASED ON THE ABOVE PREDICTION RESULTS DATAFRAME WE CAN SAY THAT LOGISTIC REGRESSION (PCA + HYPERPARAMETER TUNING) CAN BE CONSIDERED AS THE BEST MODEL AMONG ALL AS THE RECALL ON POSITIVE CLASS IS THE HIGHEST FOR THE SAME. THE TREND IN ROC_AUC OF THE MODELS ALSO JUSTIFY THE SAME.**