Group Assignment Telecom Churn

Group members:

- Truong Tuyet Lam
- Amit Kumar
- Gushpu T

Agenda

- Problem Statement & Business Goals
- Understanding and Handling Data
- Key Findings from Exploratory Data Analysis
- Model Building & Evaluation
- Conclusions and Recommendations

Problem Statement & Business Goals

Business Problem Overview

- 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.

- For many incumbent operators, retaining high profitable customers is the number one business goal.

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

- This project is based on the Indian and Southeast Asian market where pre-paid is the most common model in telecom industry.

- Definition of churn is defined based on usage. In other words, 'usage-based-churn' is applied in this case to define customers who have not done any usage, either incoming or outgoing - in terms of calls, internet etc. over a period of time.

- In the Indian and Southeast Asian markets, approximately 80% of revenue comes from the top 20% of customers (called high-value customers). Thus, if we can reduce the churn of high-value customers, we will be able to reduce significant revenue leakage.

Business Goals

A customer level data of a leading telecom firm on the Indian and South-East Asian market was provided for analysis in order to:

1. Build predictive models to identify high value customers at high risk of churn

2. Identify the important variables that are strong predictors of churn.

Understanding and Handling data

Understanding customer behavior during churn

- In churn prediction, we assume that there are three phases of the customer lifecycle :

Good phase

The customer is happy with the service and behaves as usual.

Action phase

The customer experience starts to decline.

- He/she gets a compelling offer from a competitor
- Faces unjust charges
- Becomes unhappy with service quality etc.

The customer usually shows different behavior than in the 'good' months.

> It is crucial to identify **high-churn-risk customers** in this phase, since some corrective actions can be taken at this point (such as matching the competitor's offer/improving the service quality etc.)

Churn phase

The customer is said to have churned

Mapping with current dataset

- Given four-month window of the current dataset (6,7,8,9), the first two months are the ‘good’ phase, the third month is the ‘action’ phase, and the fourth month is the ‘churn’ phase.

Good phase | Action phase | Churn phase

Month 6,7

Month 8

Month 9

Full analysis will be applied based on data of these 3 months.

- Month 9 data is used for churn tagging.
- Then all corresponding month 9 related columns will be removed.

Handling data

- A couple of steps are done to handling and cleaning data:
- Treating missing values by:
  - Removing the columns having more than 30% missing values
  - Removing date related columns given they provide no meaning in this analysis.
  - Removing the rows having more than 50% missing values in their columns
- Filtering high value customers who are defined to have recharged with an amount more than or equal to the 70th percentile of the average recharge amount in the first two months (the good phase).
- Tagging churners and delete all the attributes corresponding to the churn phase
- Treating outliers: given our datasets are quite huge, therefore, we can exclude those points below 10^{th} and above 90^{th} percentile of remaining numeric columns.
- Treating data imbalance given the ratio of churners is 5% - 10% by applying SMOTE method.

Create new usage-based columns

- In this study, since the usage-based definition is applied to define churn. We need to create some new columns to investigate the behavioral changes of churned customers in action phase:

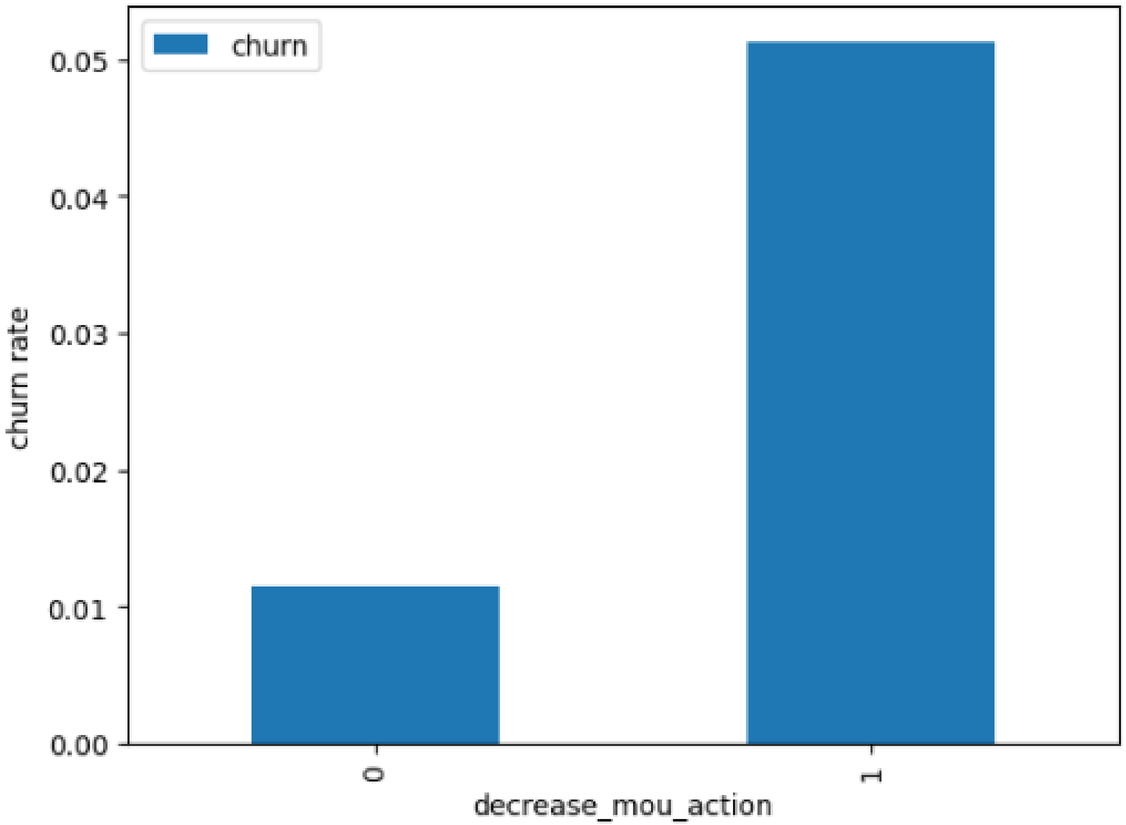
New column name	Purpose
decrease_mou_action	Indicates whether the minutes of usage of the customer (sum of incoming and outgoing calls) has decreased in the action phase than the good phase.
decrease_rech_num_action	indicates whether the frequency of recharge of the customer has decreased in the action phase than the good phase.
decrease_rech_amt_action	indicates whether the amount of recharge of the customer has decreased in the action phase than the good phase.
decrease_arpu_action	indicates whether the average revenue per customer has decreased in the action phase than the good phase.
decrease_vbc_action	indicates whether the volume based cost of the customer has decreased in the action phase than the good phase.

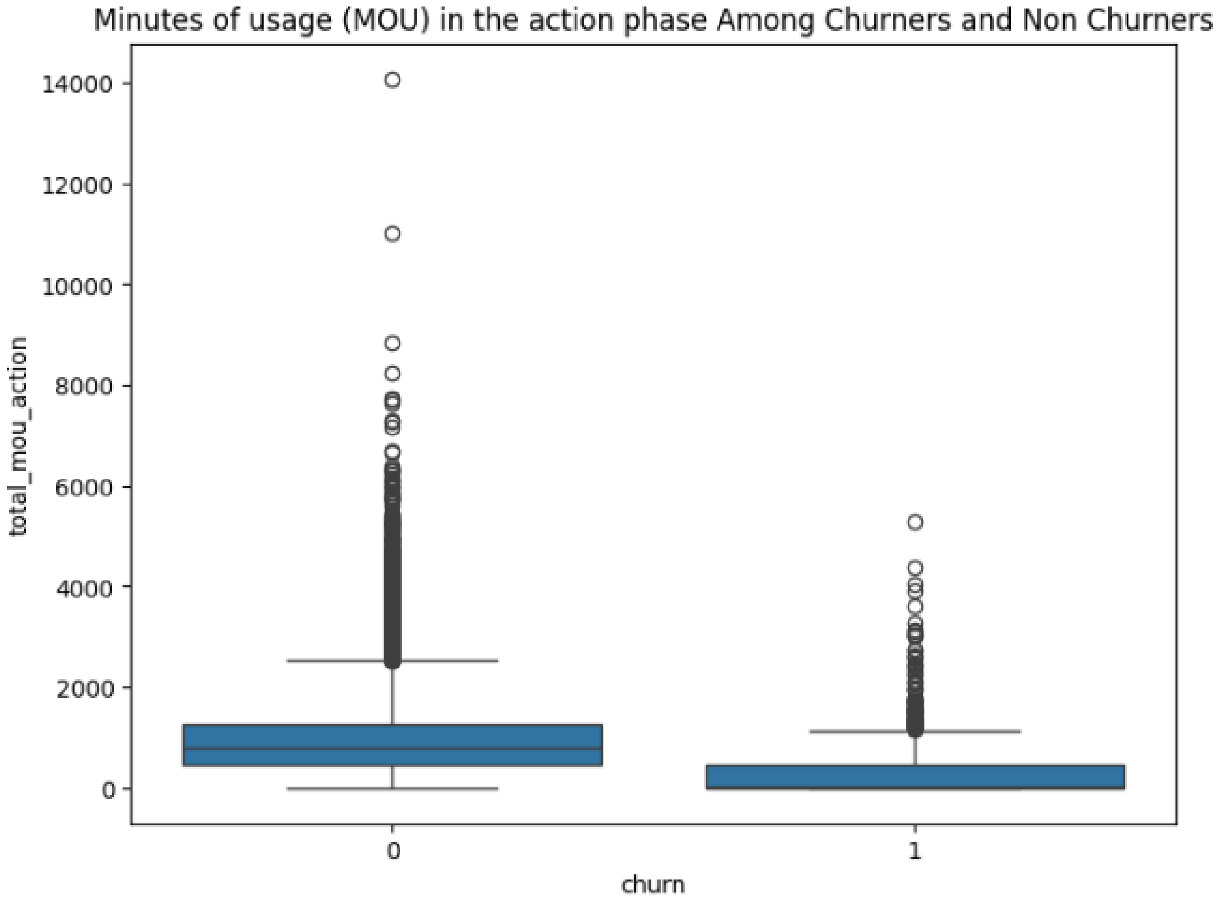
Key Findings from Exploratory Data Analysis

Minute of Usage In Action Month

- The churn rate is higher among the customers whose minutes of usage (mou) decreased in the action phase than the good phase.
- The average Minutes of usage(MOU) of the churn customers is lower than that of the non churn customers. In other words, the churners tend to reduce the MOU in action phase.

Churn rate among those customers who decreased or not decrease the MOU in action phase

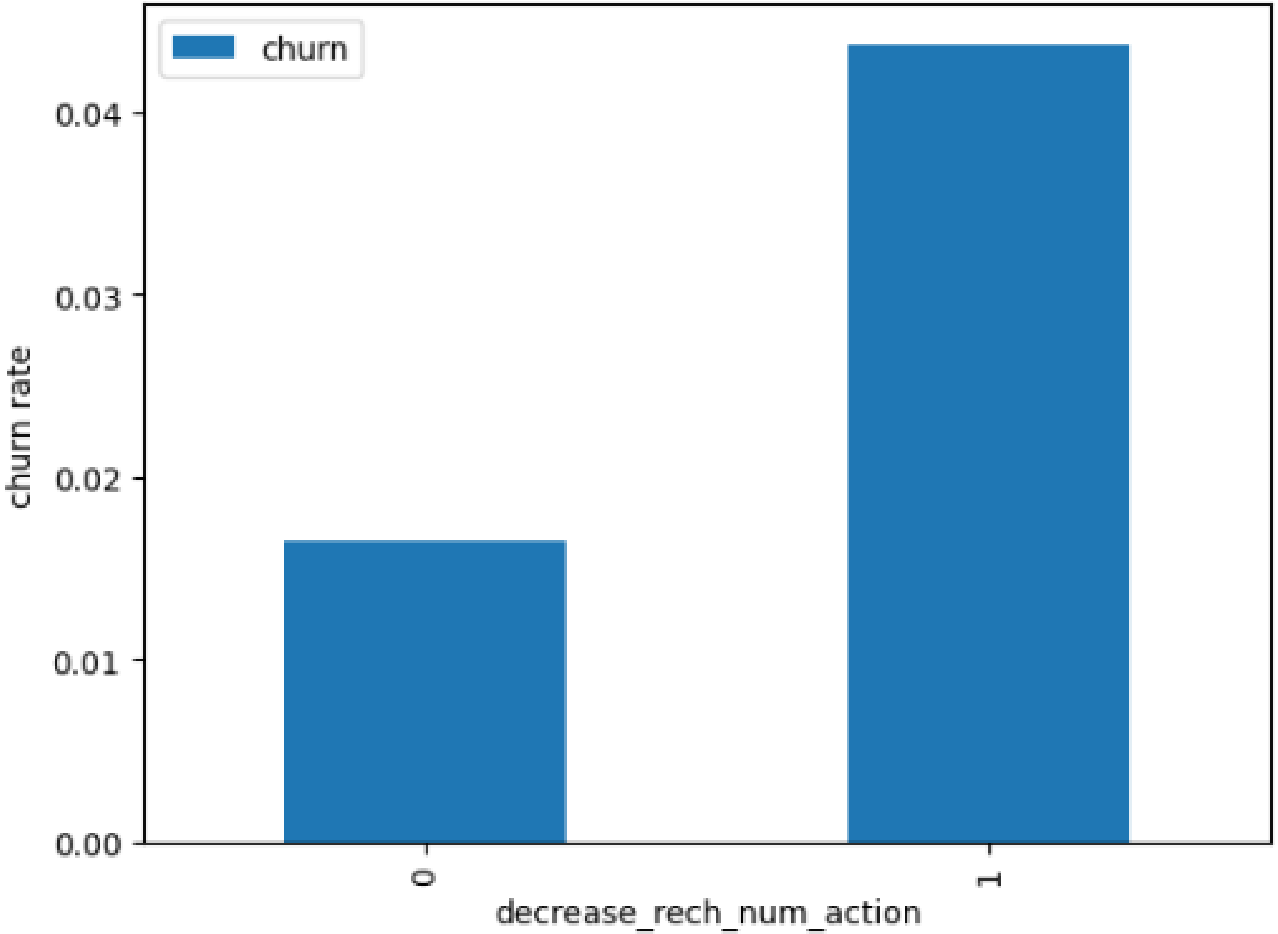




Frequency of Recharge in Action month

Similarly, the churn rate is also higher among the customers, whose frequency of recharge in the action phase is lower than the number in good phase.

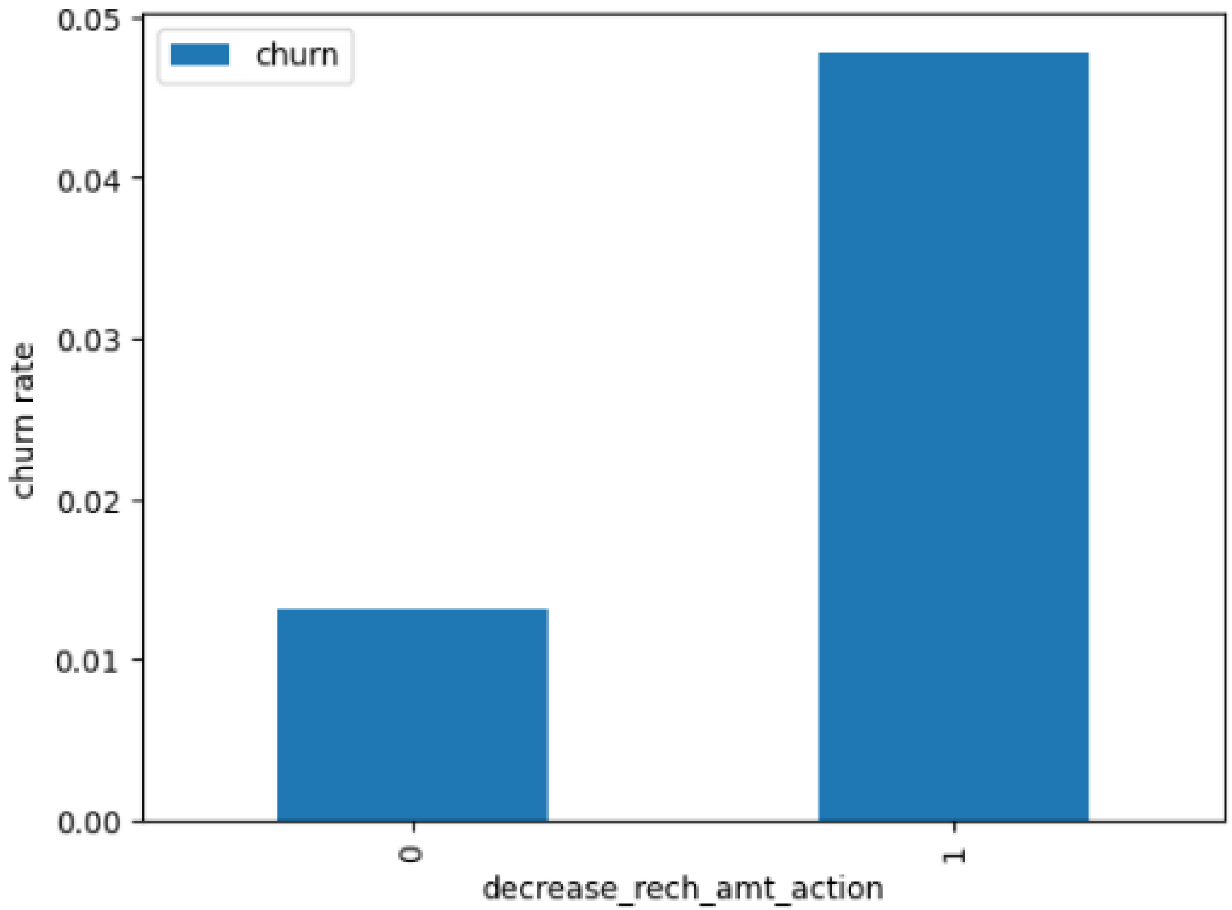
Churn rate among those customers who decreased or not decrease the recharge frequency in action phase



Total Recharge Amount In Action Month

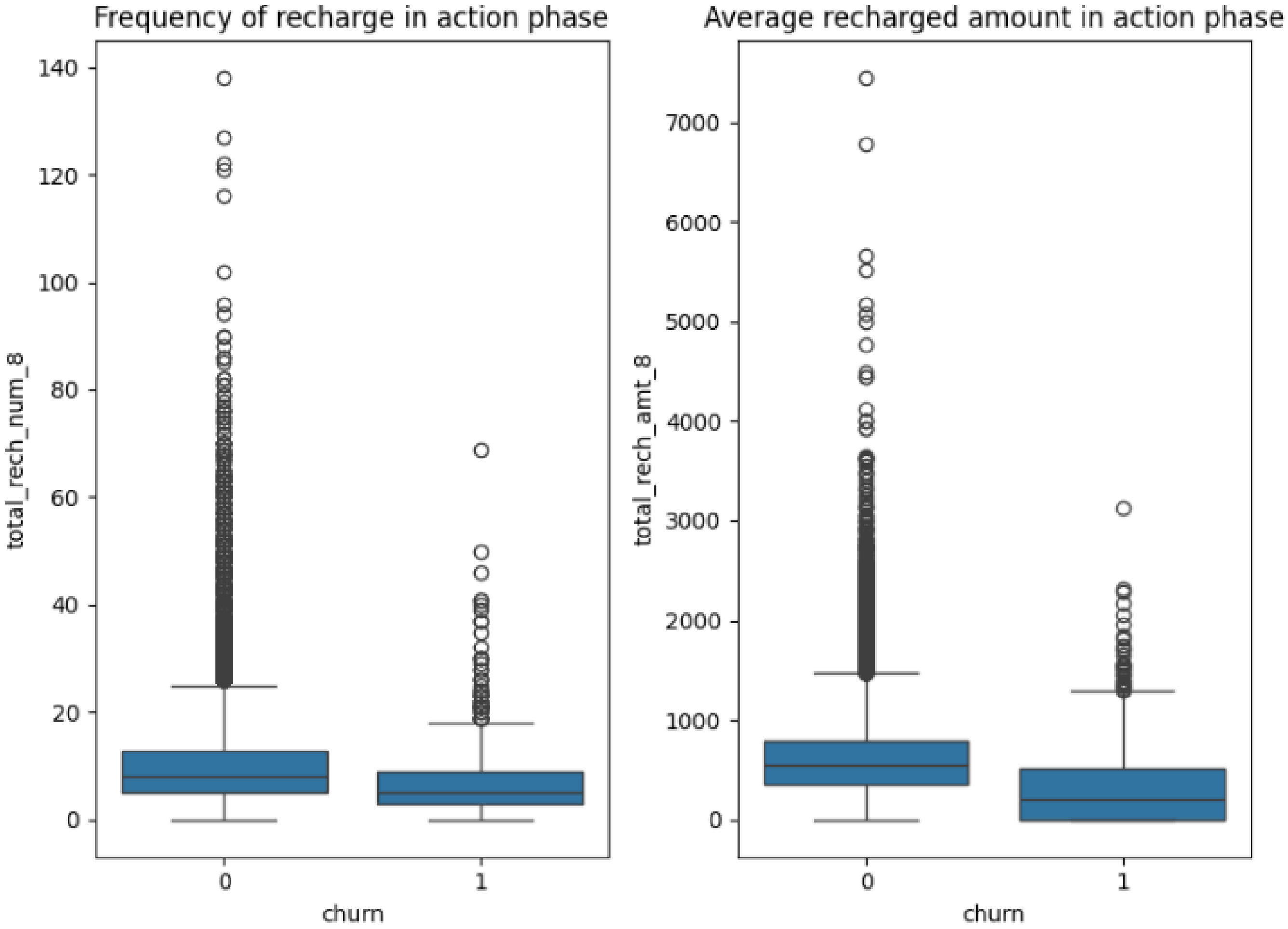
The churn rate is significantly higher among the customers, whose amount of recharge in the action phase is lower than the amount in good phase.

Churn rate among those customers who decreased or not decrease the recharge amount in action phase



Frequency and average amount of recharge in action month

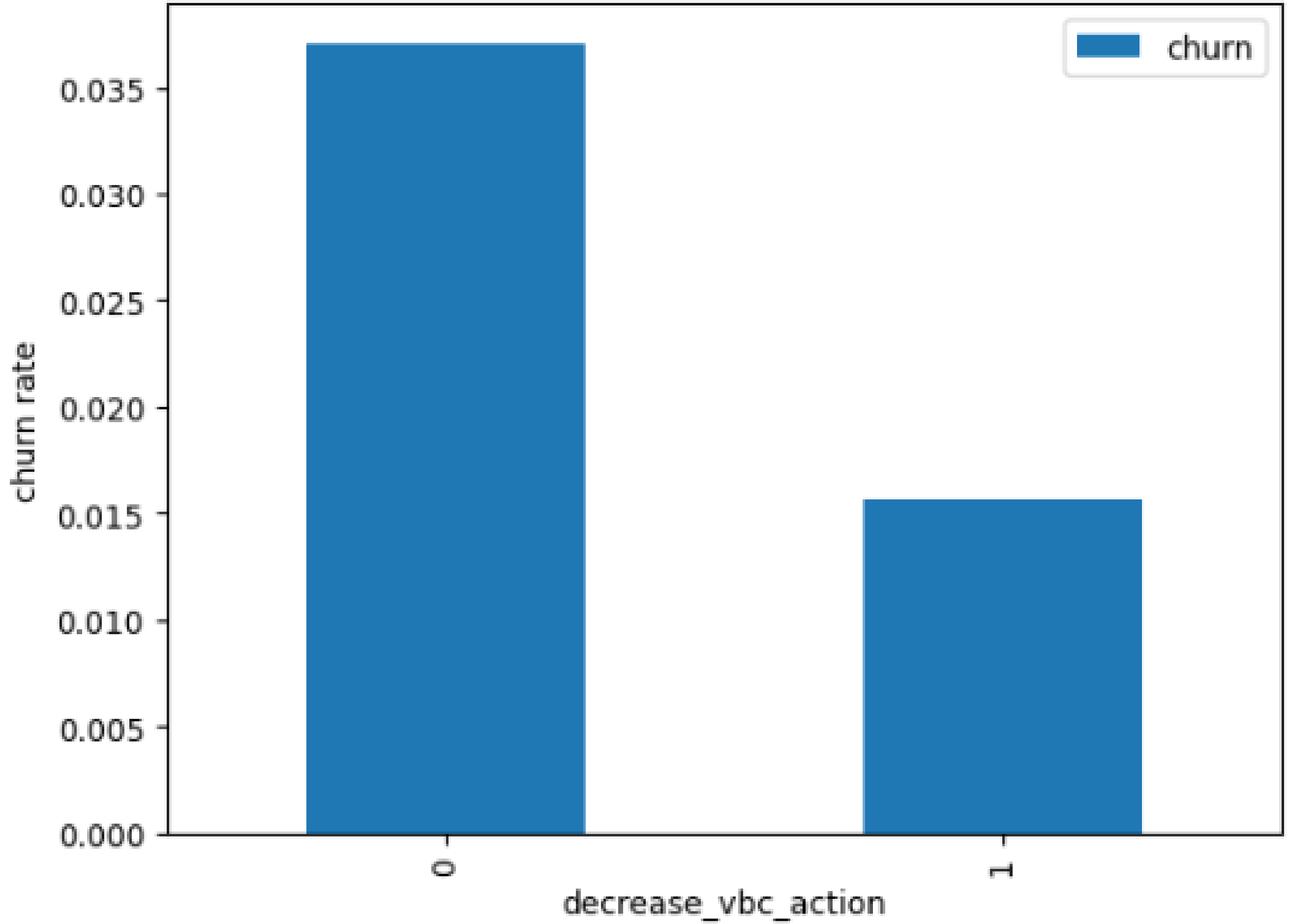
There is a decrease in both frequency and average amount recharge among churners compared to non churners.



Volume based cost in Action Month

The churn rate is much higher among the customers whose volume-based cost in action month increased. That means the customers do not do the monthly recharge more when they are in the action phase.

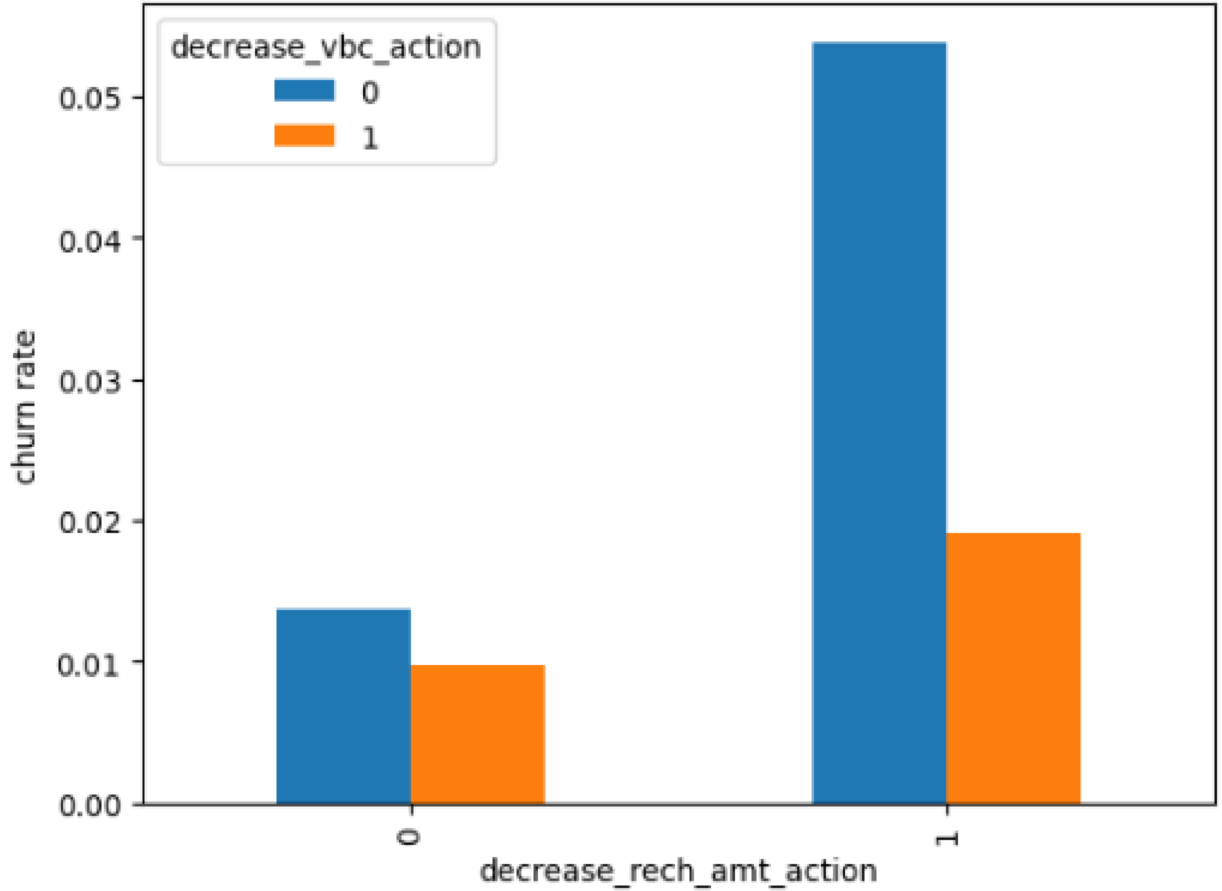
<u>Churn rate among those customers who decreased or not decrease the volume-based cost in action phase</u>



Decreasing recharge amount vs volume based cost in action phase

The churn rate is more for the customers, whose recharge amount is decreased along with the volume based cost is increased in the action month.

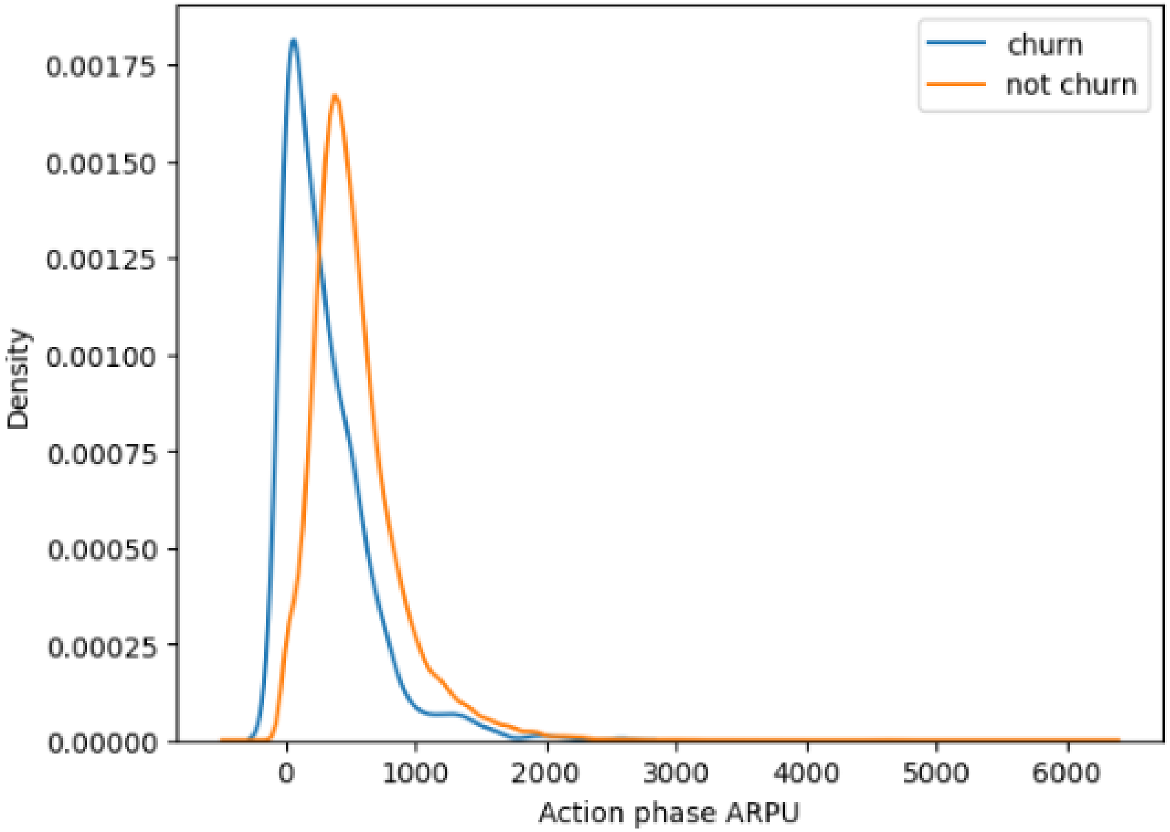
<u>Churn rate by the decreasing recharge amount and volume based cost in the action phase</u>

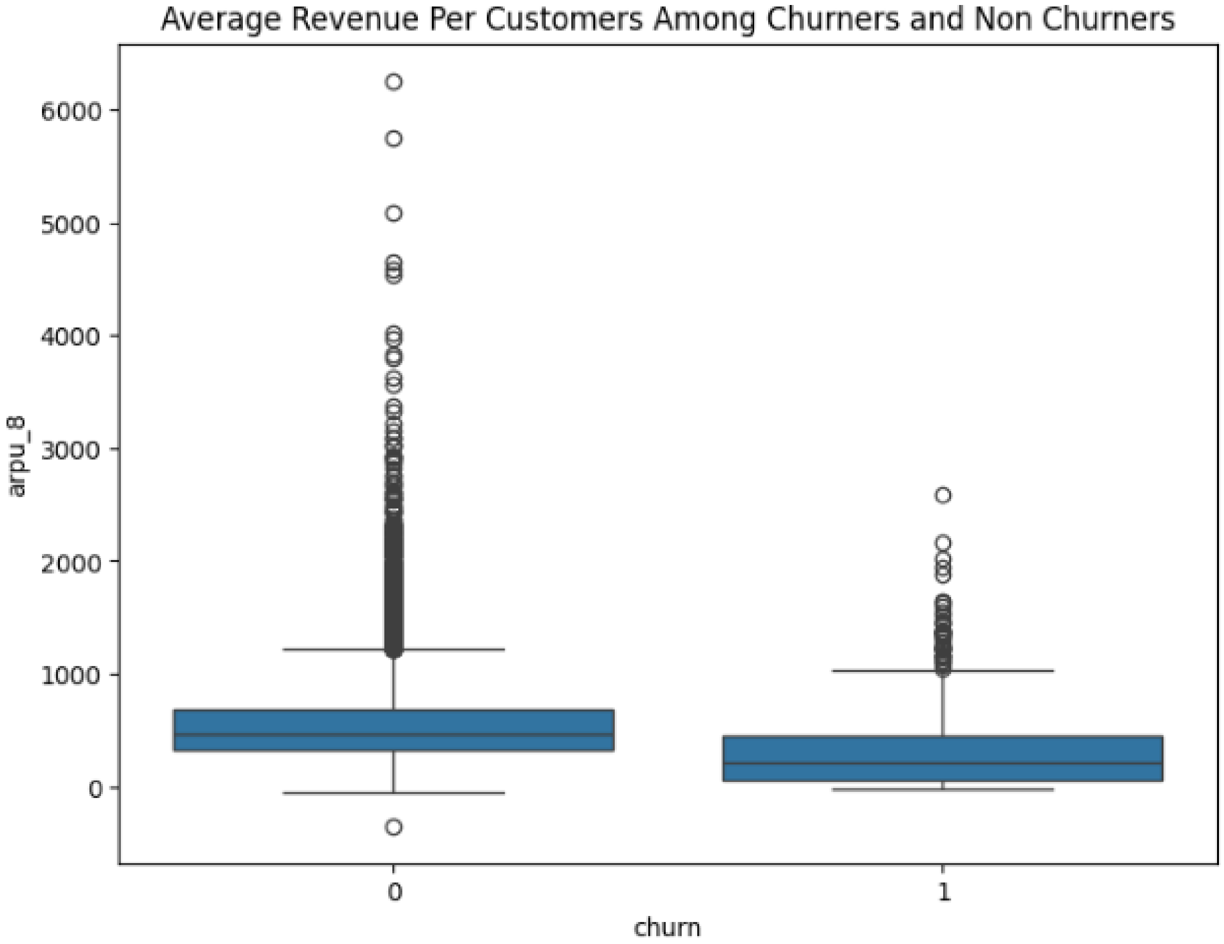


Average Revenue Per Customer

- Average revenue per user (ARPU) for the churners varies from 0 to 900 meanwhile the ARPU among non-churners is higher, varies in a range from 0 to 1000. The higher ARPU customers are less likely to be churned.
- Also, the second chart reveals the average revenue per customers among churners is lower than that in non churners.

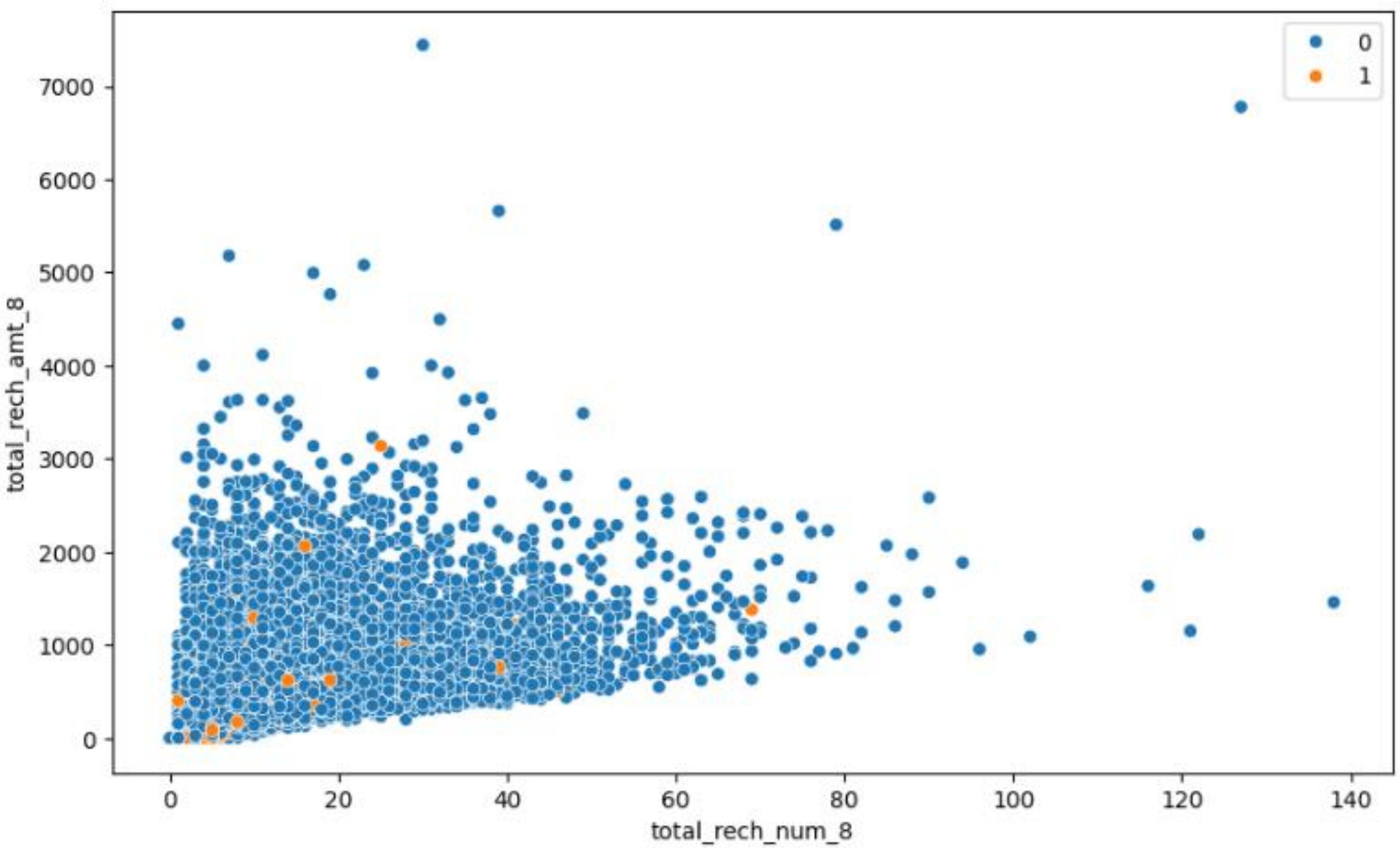
Average Revenue Per Customers between Churners and Non-Churners





Recharge amount vs. frequency of recharge in action month

The recharge number and the recharge amount are mostly propotional. More the number of recharge, more the amount of the recharge.



Model Building & Evaluation

Model Building on Train Set

After multiple steps of RFE as well as checking p-values and VIF scores, a model is generated with 13 variables as below:

Generalized Linear Model Regression Results

Dep. Variable:	churn	**No. Observations:**	42850
Model:	GLM	**Df Residuals:**	42836
Model Family:	Binomial	**Df Model:**	13
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-15365.
Date:	Mon, 10 Mar 2025	**Deviance:**	30731.
Time:	08:27:59	**Pearson chi2:**	4.61e+06
No. Iterations:	11	**Pseudo R-squ. (CS):**	0.4878
Covariance Type:	nonrobust		

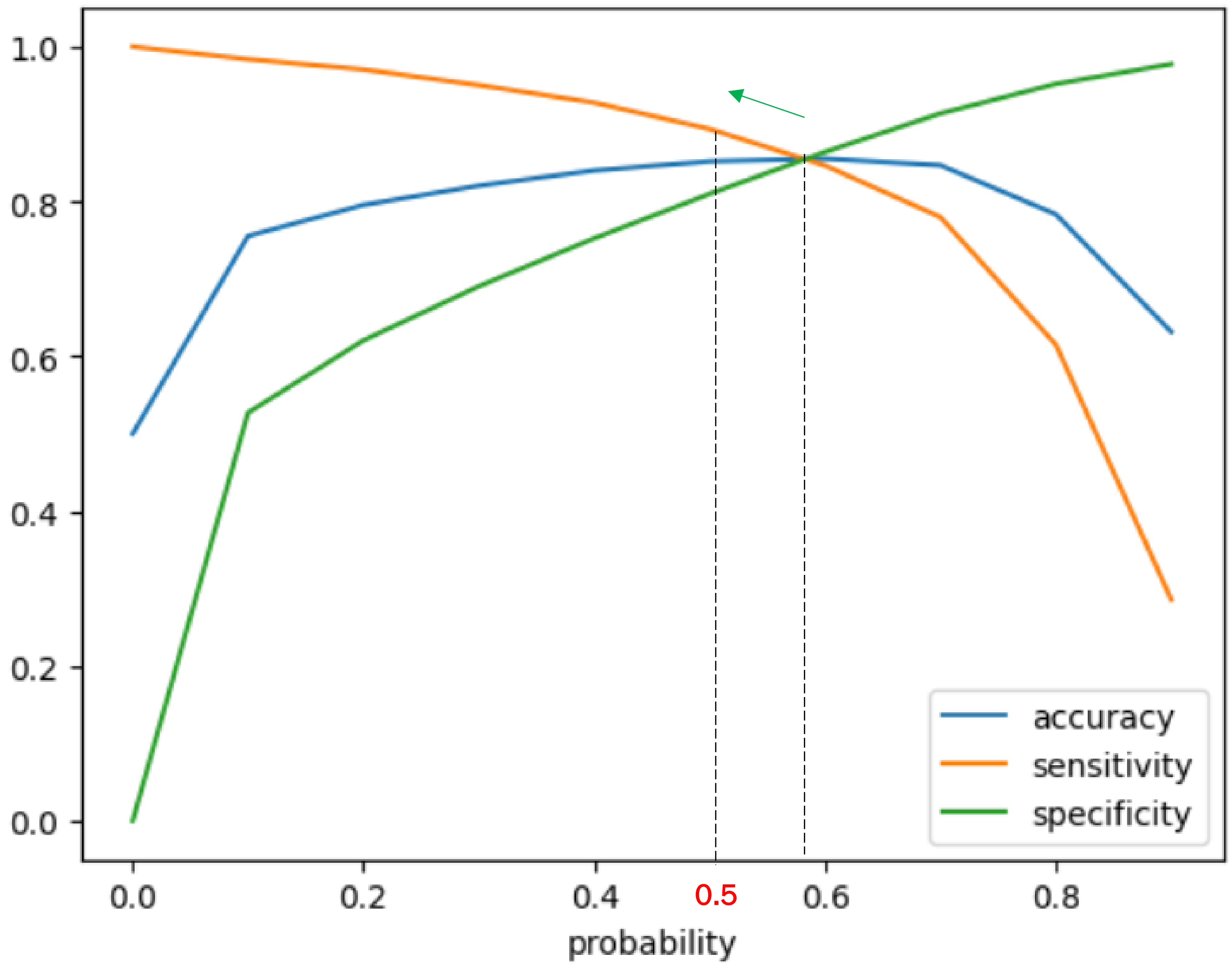
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.9762	0.030	-32.032	0.000	-1.036	-0.916
offnet_mou_7	0.4689	0.023	19.979	0.000	0.423	0.515
roam_og_mou_8	0.5967	0.024	25.325	0.000	0.551	0.643
std_og_t2t_mou_7	0.7488	0.025	30.029	0.000	0.700	0.798
std_og_t2m_mou_8	0.3167	0.033	9.723	0.000	0.253	0.380
isd_og_mou_8	-1.1466	0.193	-5.954	0.000	-1.524	-0.769
og_others_7	-1.6873	0.729	-2.315	0.021	-3.116	-0.258
total_og_mou_8	-1.2600	0.038	-32.931	0.000	-1.335	-1.185
loc_ic_mou_8	-2.9608	0.057	-51.923	0.000	-3.073	-2.849
std_ic_t2f_mou_8	-0.7640	0.074	-10.357	0.000	-0.909	-0.619
ic_others_8	-1.3762	0.124	-11.076	0.000	-1.620	-1.133
monthly_2g_8	-0.8935	0.044	-20.244	0.000	-0.980	-0.807
monthly_3g_8	-1.0030	0.046	-21.989	0.000	-1.092	-0.914
decrease_vbc_action	-1.3162	0.061	-21.446	0.000	-1.436	-1.196

	Features	VIF
3	std_og_t2m_mou_8	2.88
6	total_og_mou_8	2.86
0	offnet_mou_7	1.74
2	std_og_t2t_mou_7	1.45
7	loc_ic_mou_8	1.34
12	decrease_vbc_action	1.09
1	roam_og_mou_8	1.06
11	monthly_3g_8	1.06
10	monthly_2g_8	1.06
9	ic_others_8	1.01
8	std_ic_t2f_mou_8	1.01
4	isd_og_mou_8	1.00
5	og_others_7	1.00

Evaluating Model

The main goal is to retain the customers, who have the possibility to churn. Thus, for this case, we should pay more attention to achieve higher Sensitivity/Recall score than the accuracy. Because we need to care more about churn cases than non-churn cases. There should not be a problem, if we consider some non-churn customers as churn customers and provide them some incentives for retaining them. Hence, the sensitivity score is more important here.

Given the chart show 0.6 cut off is the optimal point, but as sensitivity is more important so let's aim to cut off at 0.5.



Model Evaluation on Train Set - Cut off at 0.5

At the cut-off point for churn probability at 0.5, the confusion matrix is quite balance among true/false or negatives/positives. Besides, as we can achieve the higher sensitivity score, then it is fine to proceed further with this model.

Confusion Matrix

Actual\ Predicted	Not_churn	Churn
Not_churn	17348	4077
Churn	2280	19145

Accuracy Score: 85%

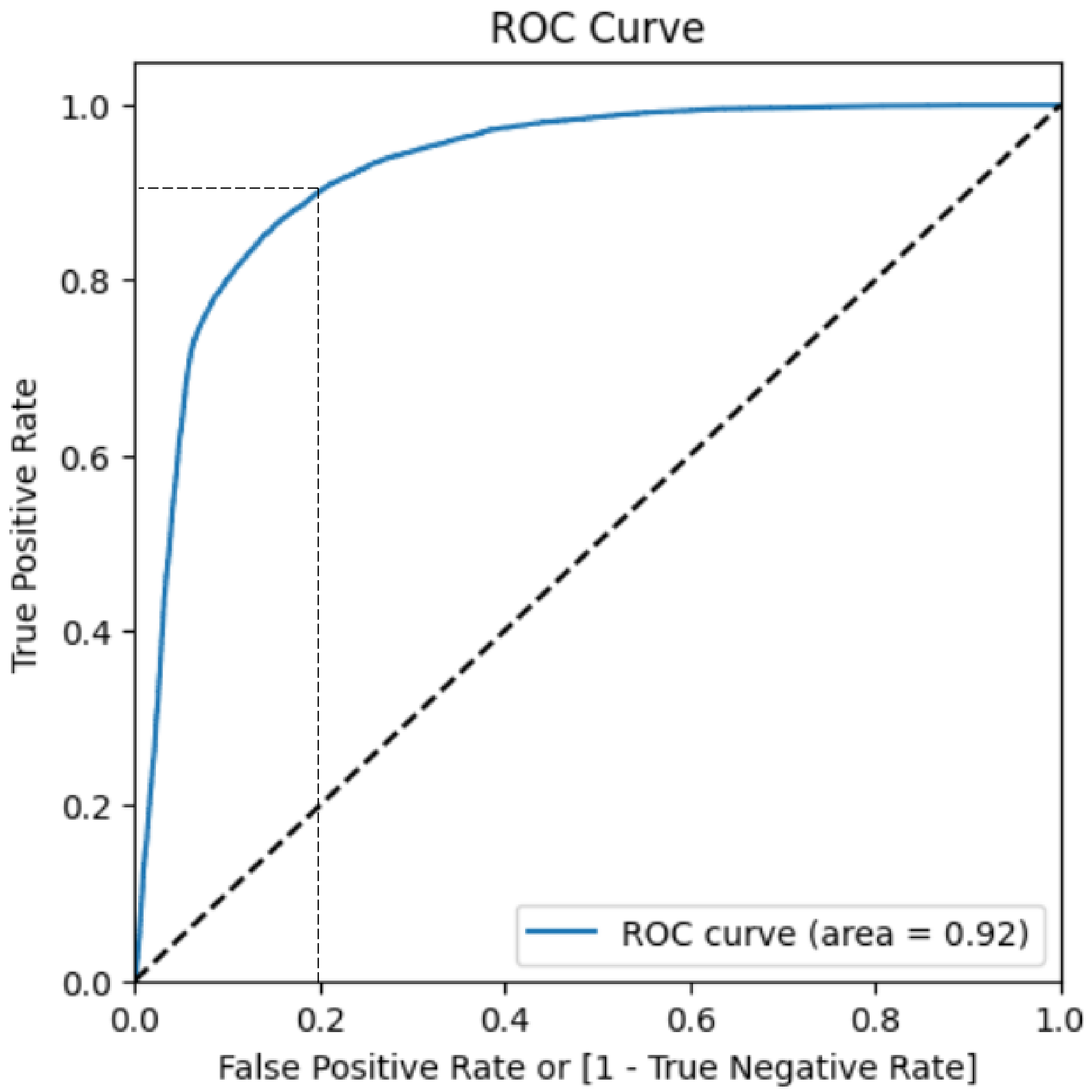
Sensitivity Score: 89%

Specificity Score: 80%

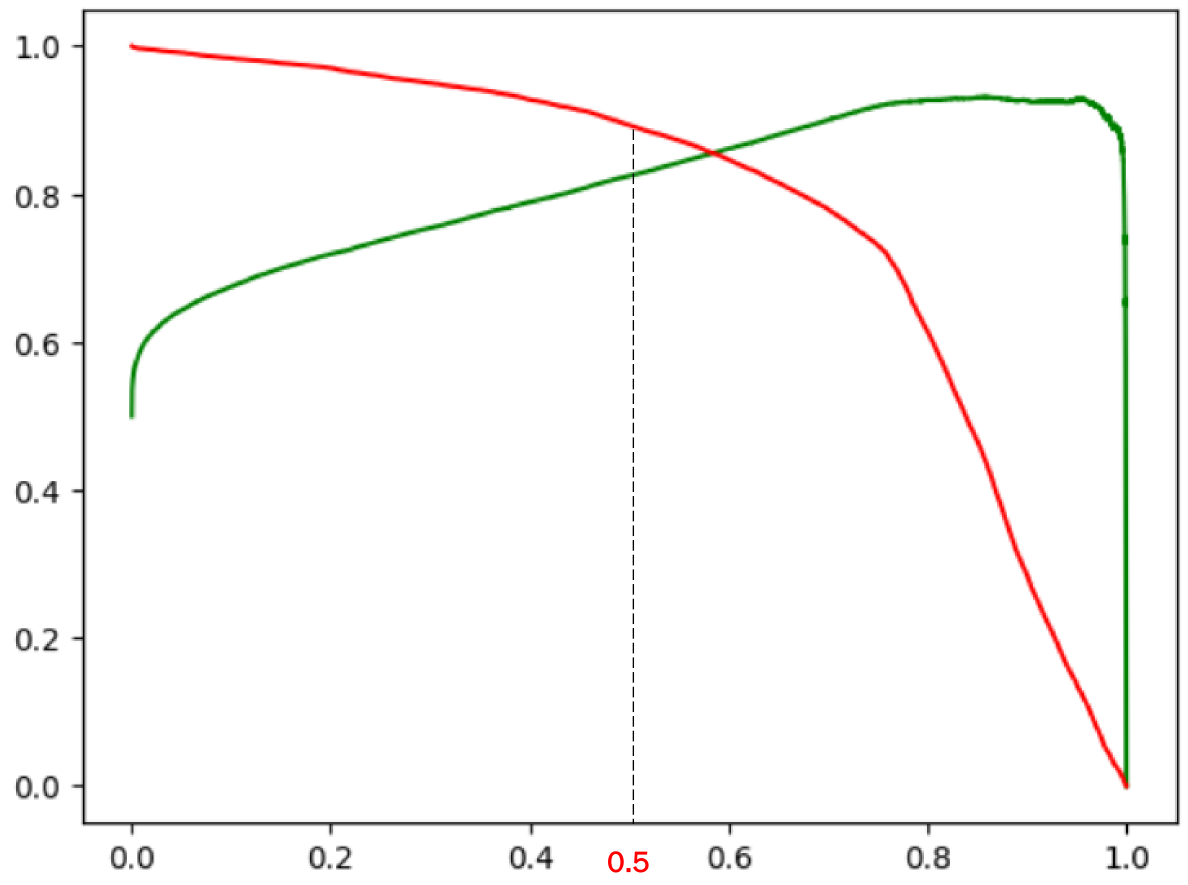
True Positive Rate: 82%

False Positive Rate: 20%

Model Evaluation on Train Set



Precision & Recall Trade Off



Precision: 82%

Recall: 82%

Prediction on Test Set - Cut off at 0.5

At the cut-off point for churn probability at 0.5, the confusion matrix of test set is not so balanced, however, it reveals a better skew to detect the churn cases and minimize the wrong churn case detection. Moreover, the accuracy, sensitivity and specificity scores are also close to one another at high level, with sensitivity score is a bit higher than accuracy and specificity scores.

- True positives: 159
- True negatives: 4286
- False positives: 1062
- False negatives: 34

Confusion Matrix

Actual\ Predicted	Not_Churn	Churn
Not_Churn	4286	1062
Churn	34	159

Accuracy Score: 80%

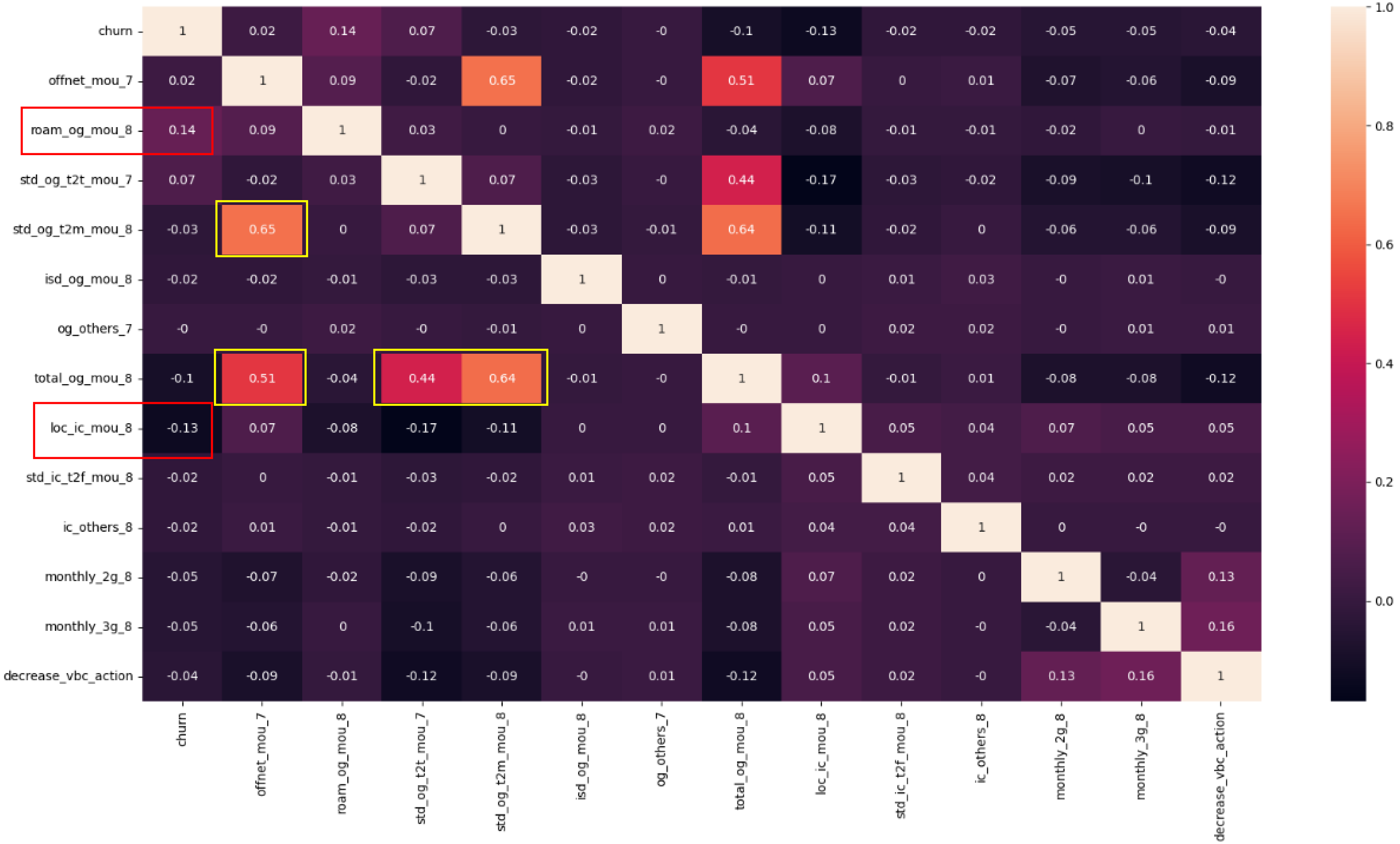
Sensitivity Score: 82%

Specificity Score: 80%

True Positive Rate: 82%

False Positive Rate: 20%

Correlation matrix



- Among top attributes, **'roam_og_mou_8'** and **'loc_ic_mou_8'** show higher correlation with churn.
- offnet_mou_7', 'std_og_t2t_mou_7' and 'std_og_t2m_mou_8' show really high correlation with 'total_og_mou_8'.

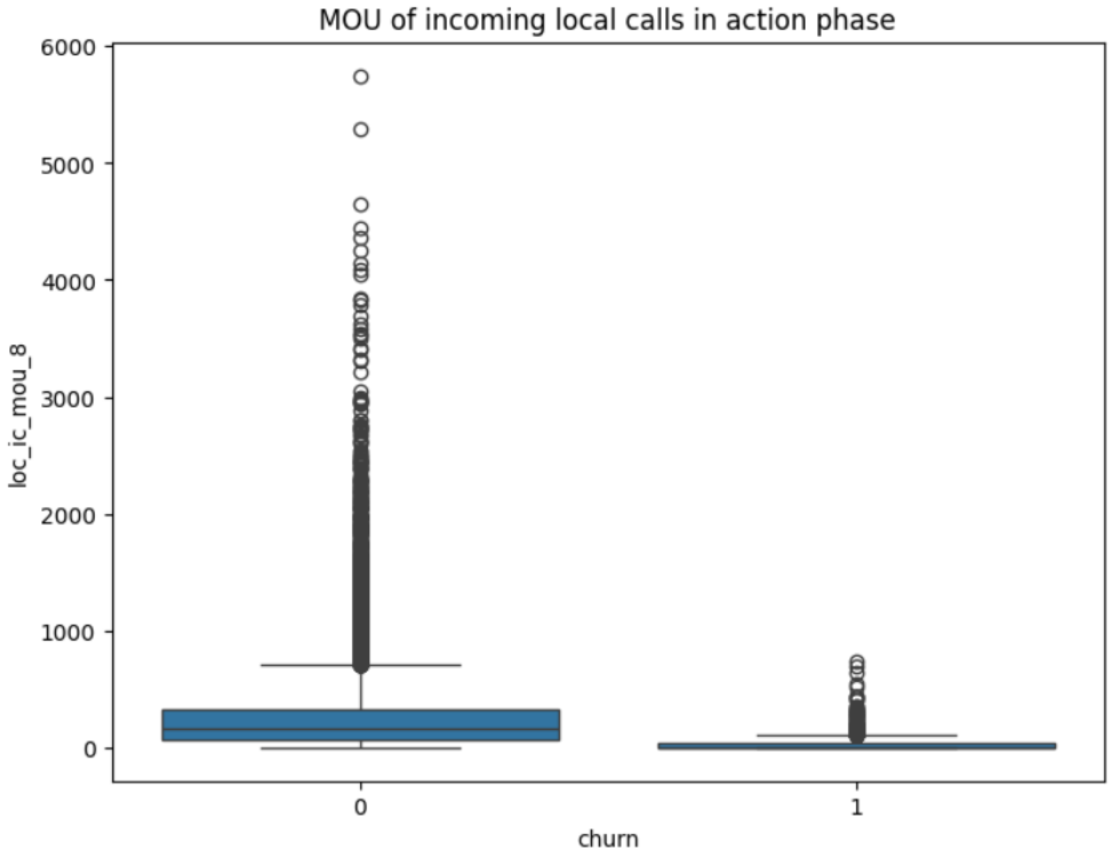
Co-efficient analysis

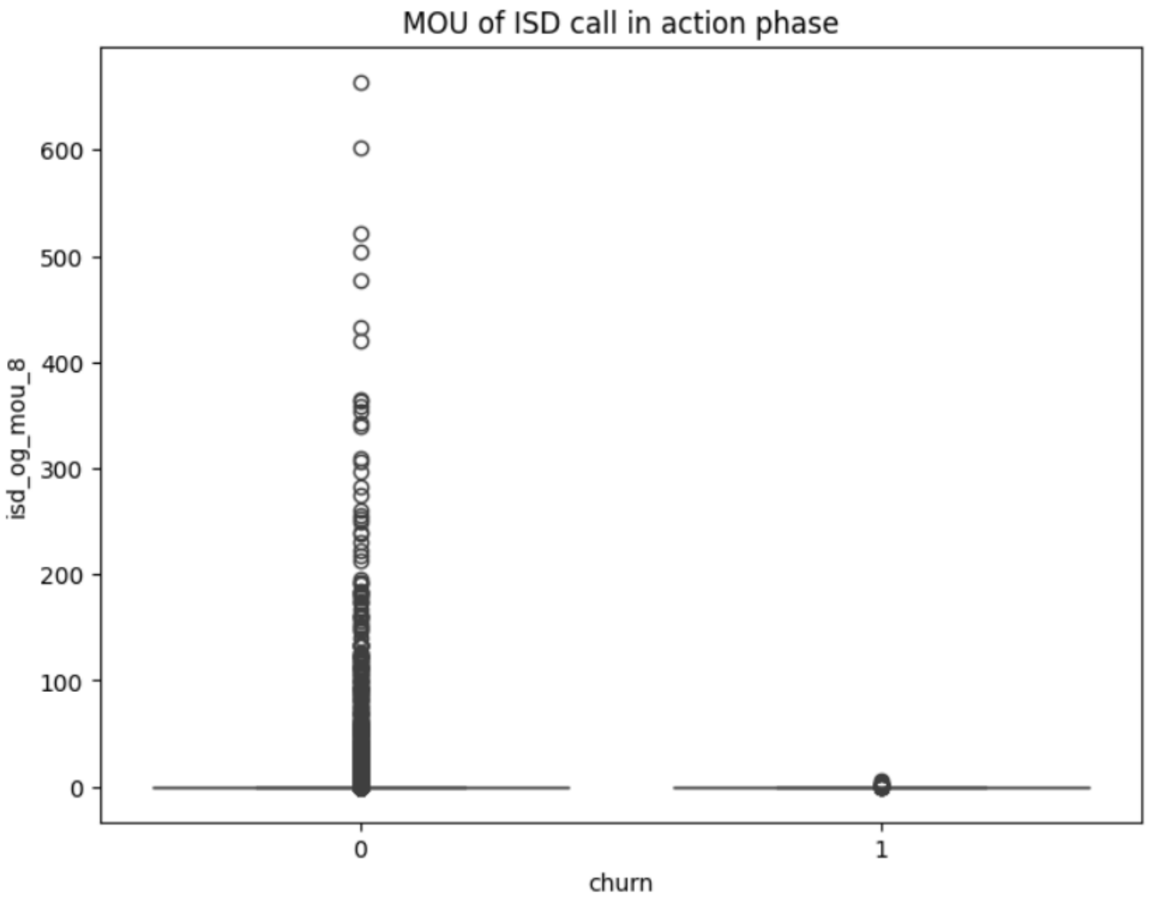
- We can see most of the top variables have negative coefficients. That means, the variables are inversely correlated with the churn probability.
- However, there are some factors having positive co-efficient on churn which we need to watch out, including:
    - ‘Offnet_mou_7’: increase in all kinds of calls outside operator T network will lead to churn.
    - ‘std_og_mou_7’ & ‘std_og_t2m_mou_8’: increase in STD call outside the calling circle may also bring some risk to churn, since this made the ‘loc_ic_mou_8’ decrease, given the strong correlation between these 3 facts.
    - ‘roaming_og_mou_8’: increase in roaming also bring some risk of churn.
- Though most of important reflecting the behaviour change in month 8 – action month, there are also factors reflecting behavior changes in month 7 – even in good phase. Thus, we need to monitor the changes in customers behaviors since good phase.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.9762	0.030	-32.032	0.000	-1.036	-0.916
offnet_mou_7	0.4689	0.023	19.979	0.000	0.423	0.515
roam_og_mou_8	0.5967	0.024	25.325	0.000	0.551	0.643
std_og_t2t_mou_7	0.7488	0.025	30.029	0.000	0.700	0.798
std_og_t2m_mou_8	0.3167	0.033	9.723	0.000	0.253	0.380
isd_og_mou_8	-1.1466	0.193	-5.954	0.000	-1.524	-0.769
og_others_7	-1.6873	0.729	-2.315	0.021	-3.116	-0.258
total_og_mou_8	-1.2600	0.038	-32.931	0.000	-1.335	-1.185
loc_ic_mou_8	-2.9608	0.057	-51.923	0.000	-3.073	-2.849
std_ic_t2f_mou_8	-0.7640	0.074	-10.357	0.000	-0.909	-0.619
ic_others_8	-1.3762	0.124	-11.076	0.000	-1.620	-1.133
monthly_2g_8	-0.8935	0.044	-20.244	0.000	-0.980	-0.807
monthly_3g_8	-1.0030	0.046	-21.989	0.000	-1.092	-0.914
decrease_vbc_action	-1.3162	0.061	-21.446	0.000	-1.436	-1.196

Illustrations on the drop of certain services implies the potential risk of churn

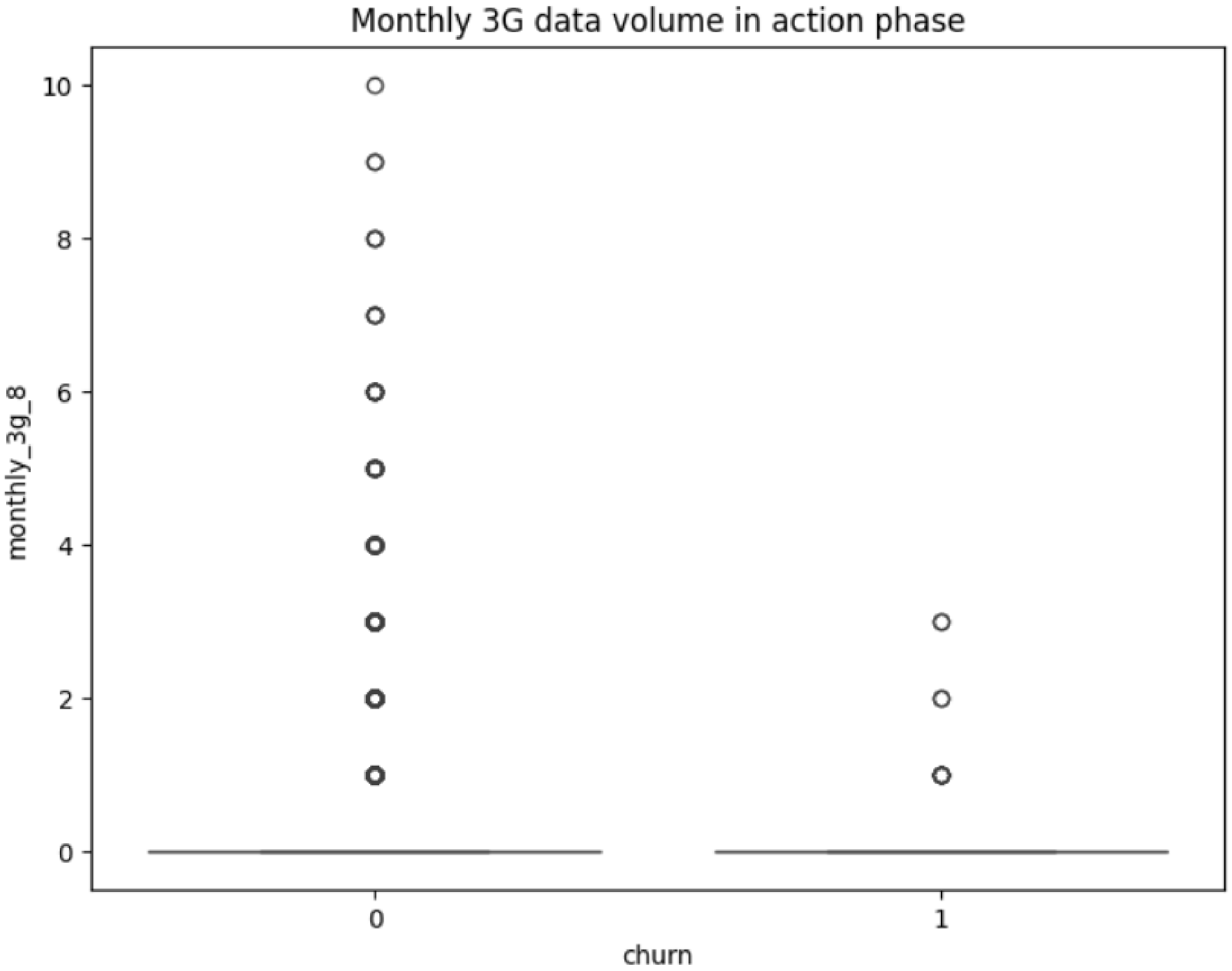
The churn customers the minutes of usage for the month of August is mostly populated on the lower side than the non churn customers. Similarly, the ISD outgoing minutes of usage for the month of August for churn customers is centered around zero.





Illustrations on the drop of certain services implies the potential risk of churn

The number of monthly 3g data for August for the churn customers are very much populated around 1, whereas of non churn customers it spreads across various numbers.



Conclusions & Recommendations

Conclusion

Based on the list of high value customers, below are top factors we need to use to monitor their behaviors.

	coef
const	-0.9762
offnet_mou_7	0.4689
roam_og_mou_8	0.5967
std_og_t2t_mou_7	0.7488
std_og_t2m_mou_8	0.3167
isd_og_mou_8	-1.1466
og_others_7	-1.6873
total_og_mou_8	-1.2600
loc_ic_mou_8	-2.9608
std_ic_t2f_mou_8	-0.7640
ic_others_8	-1.3762
monthly_2g_8	-0.8935
monthly_3g_8	-1.0030
decrease_vbc_action	-1.3162

- These key factors that the company need to watch out when observing increase in these factors

- These key factors that the company need to watch out when observing decrease in these factors

Also, even in good phase, we still need to watch out and monitor the changes in usage behaviors of customers to depict potential churn risk to have in-time corrective action plan to retain the customers.

Recommendations

- When the customers have higher demand for all calls outside of operator T network, roaming service or STD calls outside the calling circle, we should review our current price offering for these services to offer more competitive promotion/ discount to retain the customers. Otherwise, they may switch to the other operator which they tend to interact more often or provide more competitive price offering for the services they have high demand.

	coef
const	-0.9762
offnet_mou_7	0.4689
roam_og_mou_8	0.5967
std_og_t2t_mou_7	0.7488
std_og_t2m_mou_8	0.3167
isd_og_mou_8	-1.1466
og_others_7	-1.6873
total_og_mou_8	-1.2600
loc_ic_mou_8	-2.9608
std_ic_t2f_mou_8	-0.7640
ic_others_8	-1.3762
monthly_2g_8	-0.8935
monthly_3g_8	-1.0030
decrease_vbc_action	-1.3162

- On the other hands, when observing such drop in these services even from good phase to action phase, we need to have proper contact to understand why and offer better solutions for customers in order to retain them.

Thank you