


Telecom Churn Case Study

*By:
Siddharth
Vaibhav Shiradhonkar
Kurnool sai sravan*



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**.
- In this project, you will 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.

Understanding and Defining Churn

- *There are two main models of payment in the telecom industry - **postpaid** (customers pay a monthly/annual bill after using the services) and **prepaid** (customers pay/recharge with a certain amount in advance and then use the services).*
- *In the postpaid model, when customers want to switch to another operator, they usually inform the existing operator to terminate the services, and you directly know that this is an instance of churn.*
- *However, in the prepaid model, customers who want to switch to another network can simply stop using the services without any notice, and it is hard to know whether someone has actually churned or is simply not using the services temporarily (e.g. someone may be on a trip abroad for a month or two and then intend to resume using the services again).*
- *Thus, churn prediction is usually more critical (and non-trivial) for prepaid customers, and the term 'churn' should be defined carefully. Also, prepaid is the most common model in India and Southeast Asia, while postpaid is more common in Europe in North America.*
- *This project is based on the Indian and Southeast Asian market.*

Steps Followed

1. *Reading, understanding and visualising the data*
2. *Preparing the data for modelling*
3. *Building the model*
4. *Evaluate the model*

*Reading, Understanding and Visualising
the Data*

- *Importing DataSet*
- *Reading And Understanding the Dataset*
 - ◆ *Data Shape - (99999,226)*
 - ◆ *Data info*
 - ◆ *Data describe*
- *Handling Missing Data*
- *Eliminating the date columns as the date columns are not required in our analysis.*
 - ◆ *Dropping circle_id columns as the column has only 1 unique value.*
- *Filtering High value Customers*
 - ◆ *Creating column avg_rech_amt_6_7 by summing up total recharge amount of month 6 and 7. Then taking the average of the sum.*
 - ◆ *Obtaining the 70th percentile of the avg_rech_amt_6_7*
 - ◆ *Customers who have recharged more than or equal to 70 percentile, Filtering that customers.*

Data Shape = (30011, 178)

→ *Handling the Missing Values in Rows*

- ◆ *Here we are checking various missing values in the rows and removing the rows accordingly to optimise the data for evaluation.(i.e rows having more than 50% missing values.)*

→ *Tag Churners*

- ◆ *Presently label the beat clients (churn=1, else 0) in light of the fourth month as follows: The people who have not settled on any decisions (either approaching or friendly) AND have not utilized portable web even once in the stir stage.*
- ◆ *Eliminating all the attributes corresponding to the phase in churn*
- ◆ *Checking Churn Percentage = 3.39*

→ *Outliers Treatment*

- ◆ *In the separated dataset aside from mobile_number and stir segments every one of the sections are numeric sorts. Subsequently, changing over mobile_number and stir datatype to protest.*
- ◆ *Removing the outliers which are below 10th and are above 90th percentile*
- ◆ *Data Shape= (27705, 136)*

→ *Taking in New features*

◆ *Deriving new column*

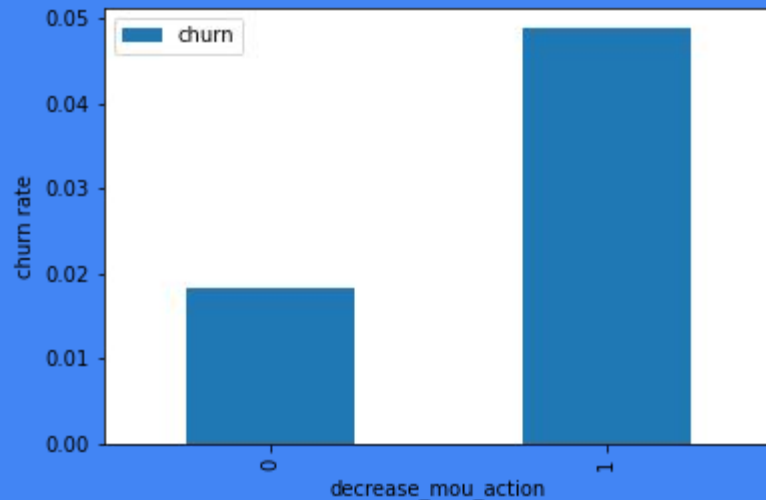
- *decrease_mou_action*
- *decrease_rech_num_action*
- *decrease_rech_amt_action*
- *decrease_arpu_action*
- *decrease_vbc_action*

Exploratory Data Analysis (EDA)

Univariate Analysis

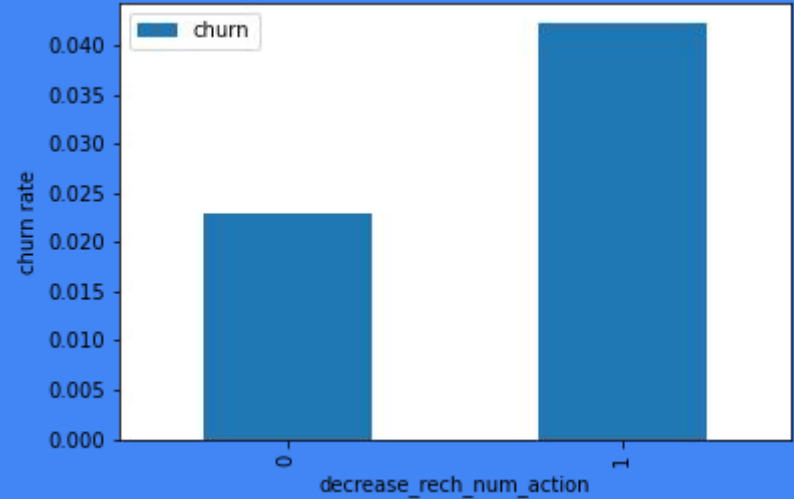
On the basis on the churn rate whether the customer decreased his MOU in action month

We can obviously see that the churn rate is something else for the clients, whose minutes of usage(mou) diminished in the activity stage than the great stage.



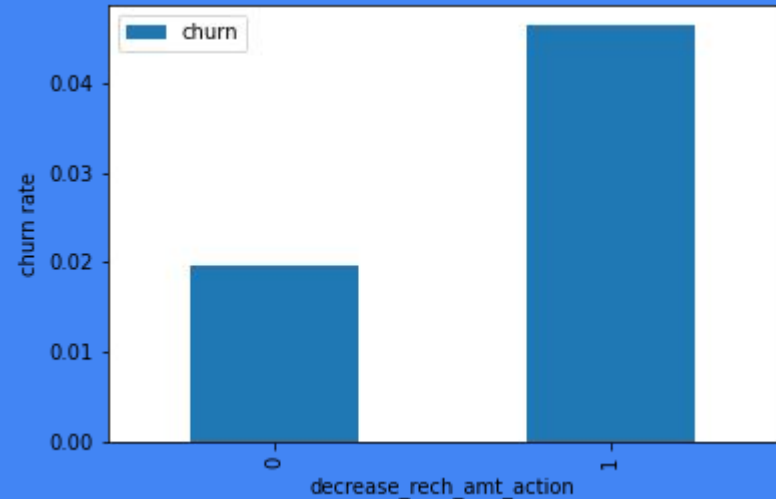
Churn rate is based on whether the customer decreased his number of recharge in the action month

True to form, the churn rate is something else for the clients, whose number of re-energize in the activity stage is lesser than the number in great stage



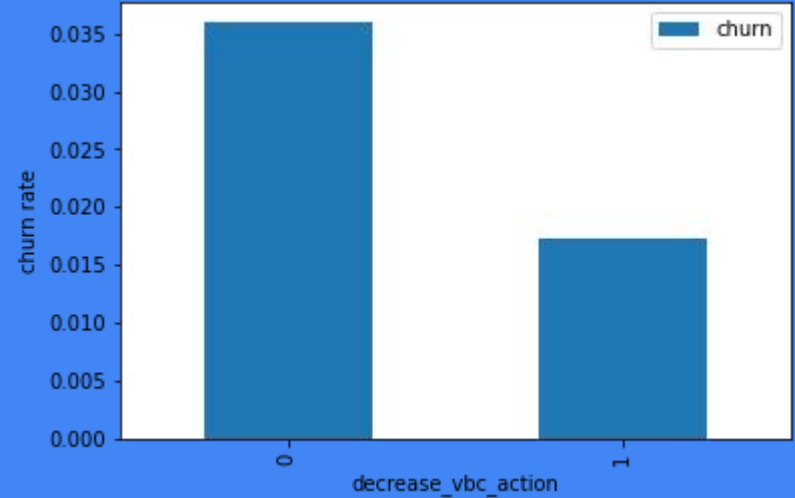
Churn rate is based on whether the customer decreased his amount of recharge in action month.

Here additionally we can plainly see that a similar way of behaving. The stir rate is something else for the clients, whose measure of re-energize in the activity stage is lesser than the sum in great stage



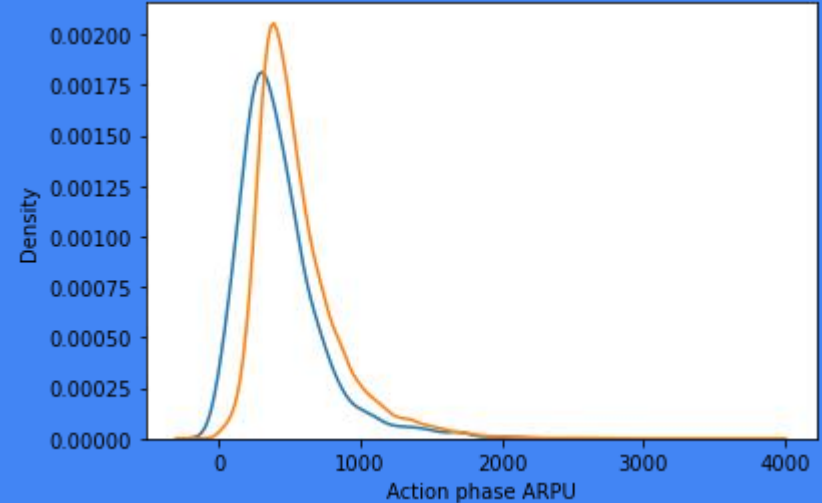
Churn rate is based on whether the customer decreased his volume based cost in action month.

The churn rate is something else for the clients, whose volume based cost in real life month is expanded. That implies the clients don't do the month to month re-energize more when they are in the activity stage. Here we can see the normal outcome.



Analysis is based on average revenue per customer of the typical income per client (churn and not churn) in the action phase.

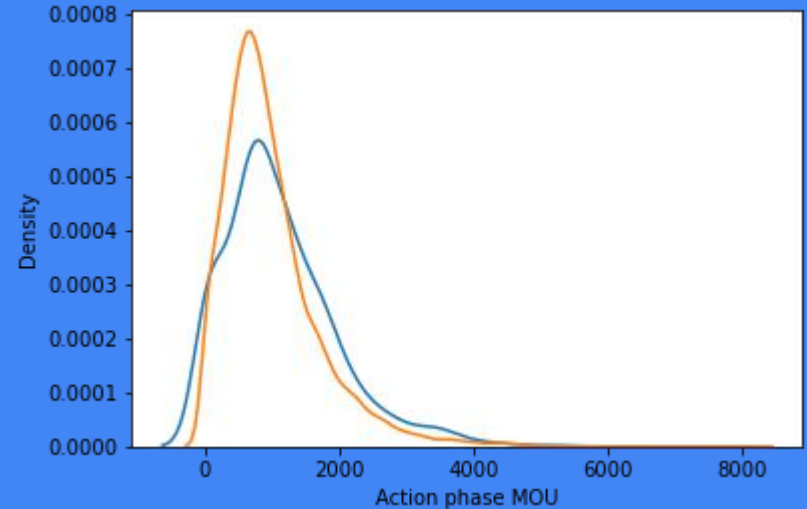
*Average revenue per user for the churned customers is mostly densed on the 0 to 900.
The higher ARPU customers are less likely to be churned.
Average revenue per user for the not churned customers is mostly densed on the 0 to 1000.*



Examination of the minutes of use MOU (churn and not churn) in the action phase

Examination of the minutes of use MOU (churn and not churn) in the activity stage

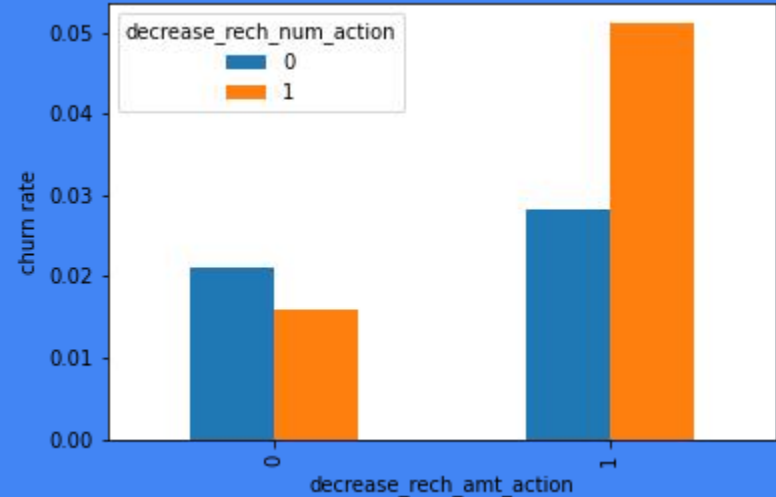
Minutes of usage(MOU) of the churn clients is for the most part populated on the 0 to 2500 territory. Higher the MOU, lesser the churn Probability.



Bivariate analysis

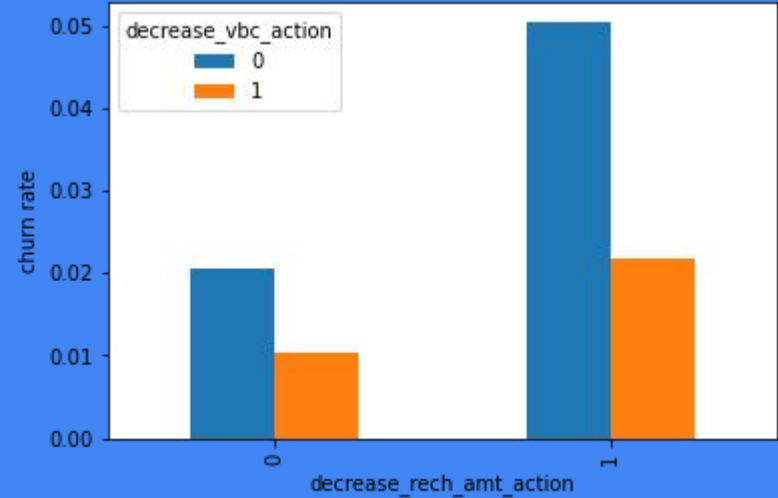
Analysis based on churn rate by the decreasing recharge amount and no of recharge in the action phase.

The churn rate is more for the customers, whose recharge amount as well as number of recharge have decreased in the action phase than the good phase.



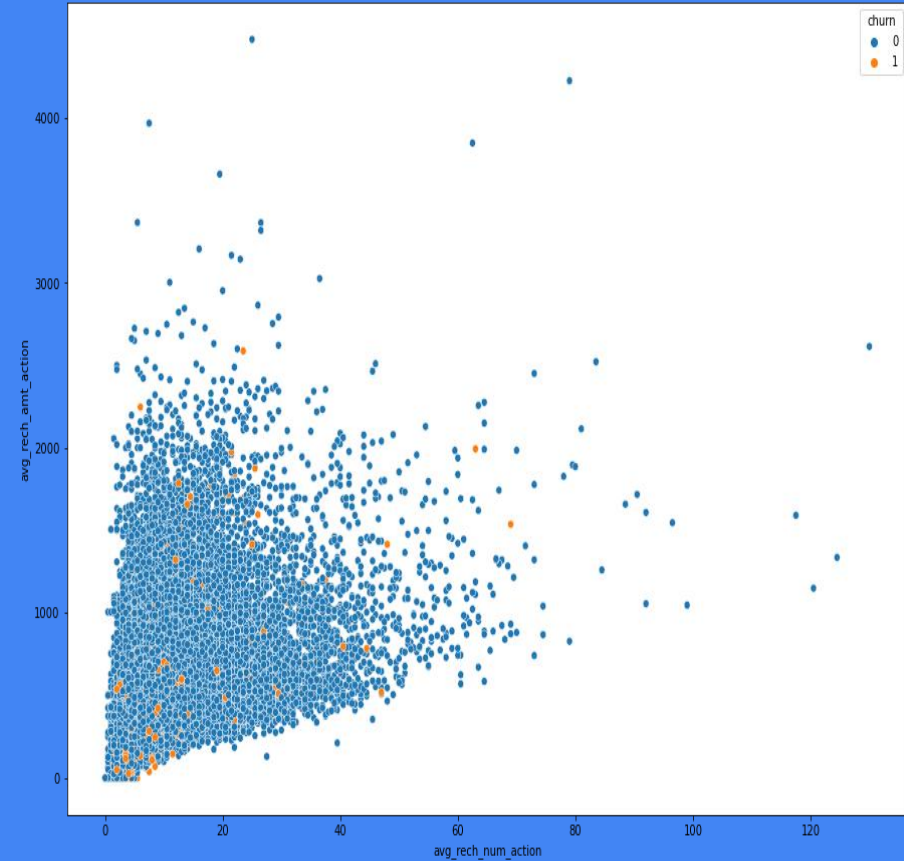
Analysis based on churn rate by the decreasing recharge amount and volume based cost in the action phase.

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.



*Analysis based on recharge amount
and no of recharge in action month.*

*Recharge number and the recharge
amount are directly propotional.*



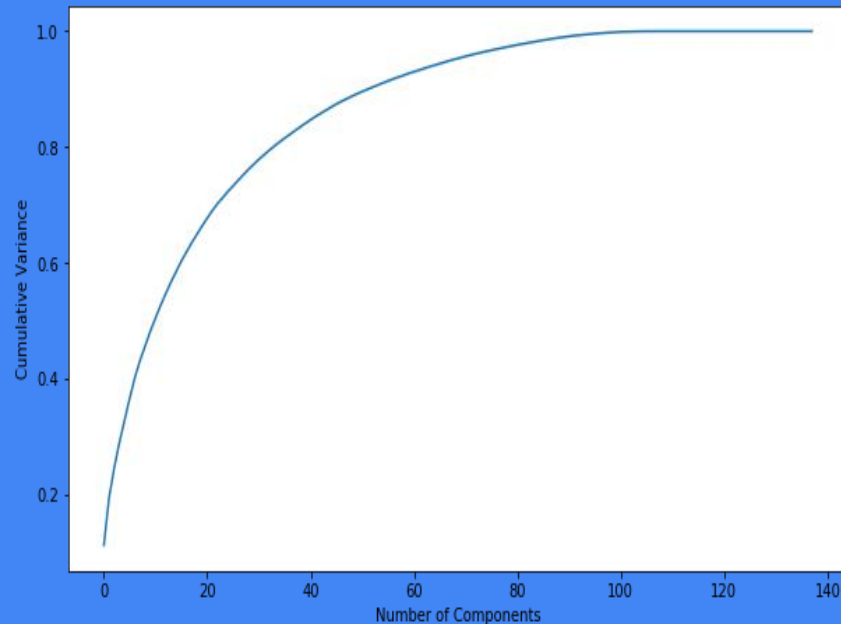
Train-Test Split

Dealing with data imbalance

We are making engineered tests by doing upsampling utilizing MOTE(Synthetic Minority Oversampling Technique).

PCA Model

We can see that 60 components make sense of almost over 90% difference of the information. Along these lines, we will perform PCA with 60 components.



Applying transformation on the test set

We are only doing Transform in the test set not the Fit-Transform. Because the Fitting is already done on the train set. So, we just have to do the transformation with the already fitted data on the train set.

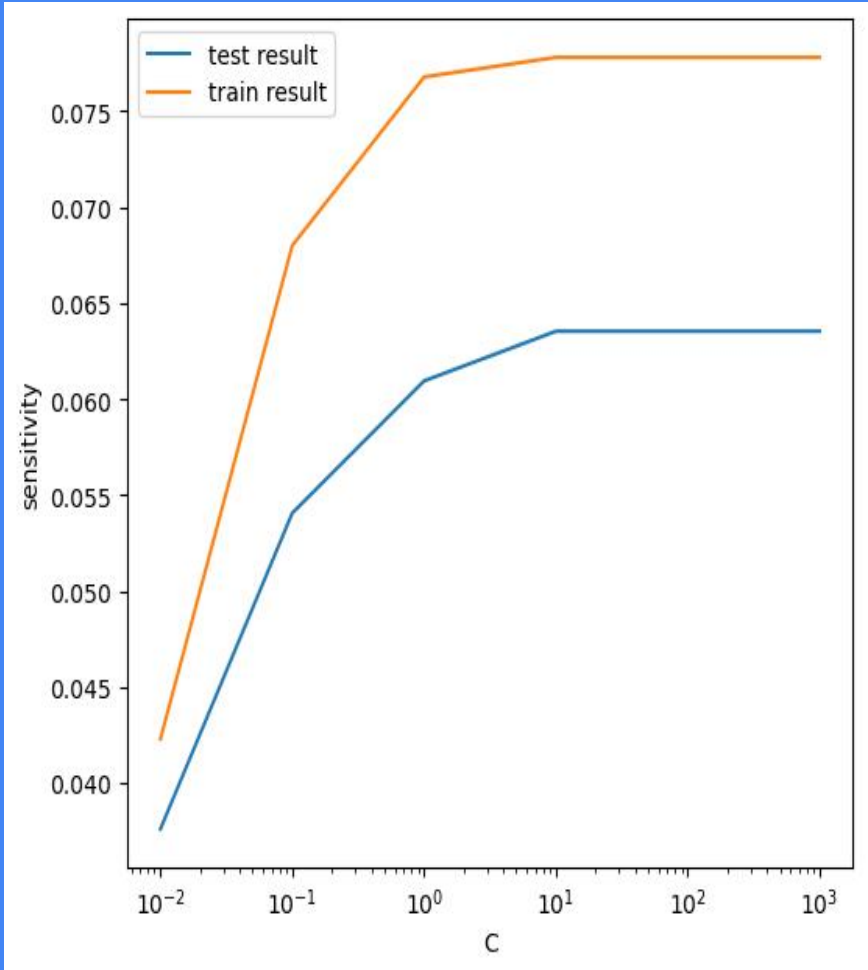
Emphasize Sensitivity/Recall than Accuracy

We are more centered around higher Responsiveness/Review score than the exactness.

Because we really want to think often more about beat cases than the not stir cases. The primary objective is to retain the clients, who have the possibility to stir. There ought not be an issue, in the event that we think about not many not stir clients as stir clients and give them a few motivating forces to holding them. Thus, the awareness score is more significant here.

Logistic regression with PCA

The Largest test sensitivity is
0.06353952242655339 at $C = 10$



Model summary

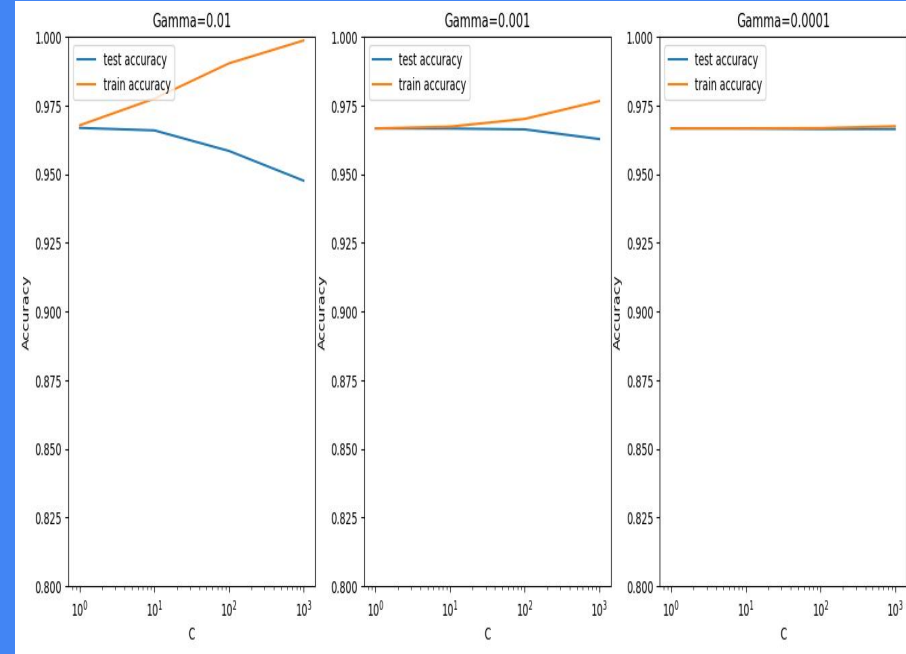
Train set

- *Accuracy = 0.966*
- *Sensitivity = 0.071*
- *Specificity = 0.997*
- *Test set*
- *Accuracy = 0.965*
- *Sensitivity = 0.088*
- *Specificity = 0.997*

Overall, the model is performing well in the test set, what it had learnt from the train set.

Support Vector Machine(SVM) with PCA

Plotting the accuracy with various C and gamma values



*The best test score is 0.9668381158635625
corresponding to hyperparameters {'C': 1, 'gamma': 0.01}*

*The higher value of gamma leads to overfitting the
model*

*we have train and test accuracy almost same, As
C=100 we have a good accuracy*

*High gamma (high non-linearity) and average value
of C*

*Low gamma (less non-linearity) and high value of
C.*

Build the model with optimal hyperparameters

Model summary*

Train set

- Accuracy = 0.966
- Sensitivity = 0.002
- Specificity = 1.0

Test set

- Accuracy = 0.965
- Sensitivity = 0.00
- Specificity = 1.0

Decision tree with PCA

Sensitivity has been decreased while evaluating the model on the test set. The accuracy and specificity is good in the test set.

Model summary*

- Train set
 - Accuracy = 0.968
 - Sensitivity = 0.179
 - Specificity = 0.99
- Test set
 - Accuracy = 0.962
 - Sensitivity = 0.093
 - Specificity = 0.993

Random forest with PCA

Model summary*

- Train set
 - Accuracy = 0.966
 - Sensitivity = 0.00
 - Specificity = 1.0
- Test set
 - Accuracy = 0.965
 - Sensitivity = 0.00
 - Specificity = 1.0

Final conclusion with PCA

In the wake of attempting a few models we can see that for achieving the best responsiveness, which was our definitive objective, the exemplary Strategic relapse or the SVM models performs well. For both the models the responsiveness was approx 96.5%. Likewise we have great precision of approx 96.5%.

Without PCA

Model analysis

There are few features have positive coefficients and few have negative.

Many features have higher p-values and hence became insignificant in the model.

Feature Selection Using RFE

Model-3

Presently from the model outline and the VIF list we can see that every one of the factors are huge and there is no multicollinearity among the factors.

Consequently, we can conclude that *Model-3 no_pca_3 will be the last model*.

Model-1 with RFE selected columns

Eliminating segment og_others_8, which is insignificant as it has the most noteworthy p-esteem 0.99.

Model-2

As we can see from the model rundown that every one of the factors p-values are huge and offnet_mou_7section has the most elevated VIF 51.29. Subsequently, erasing offnet_mou_7.

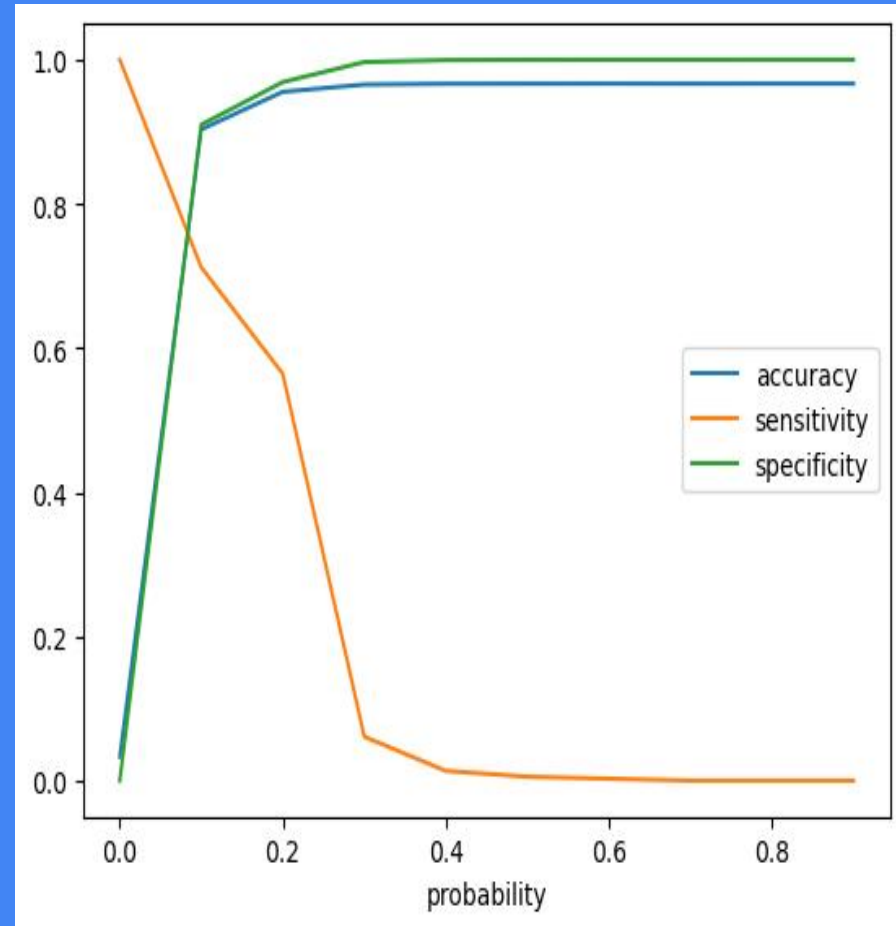
Model performance on the train set

OBSERVATIONS;

Accuracy - Becomes stable around 0.6

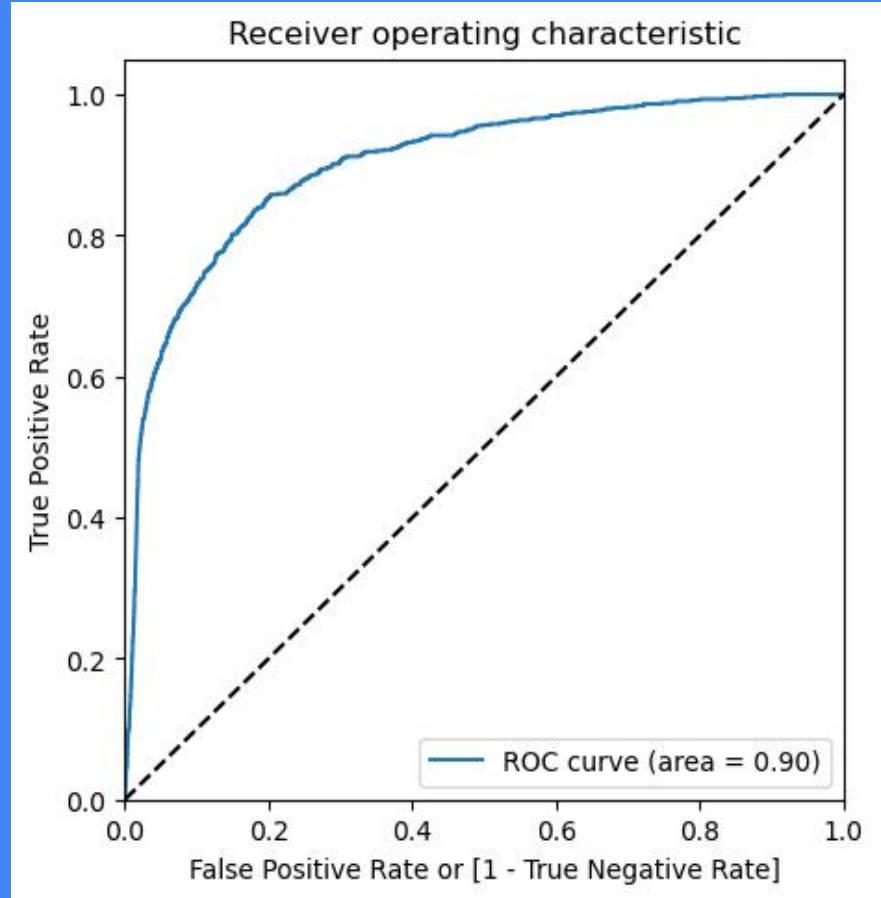
Sensitivity - Decreases with the increased probability.

*Specificity - Increases with the increased probability.
we are taking 0.5 instead of 0.6 for achieving higher sensitivity, which is our main aim.*



Plotting the ROC Curve (Trade off between sensitivity & specificity)

We can see the region of the ROC curve is closer to 1, which is the Gini of the model.



Testing the model on the test set

Model summary

- Train set
 - Accuracy:- 0.966
 - Sensitivity:- 0.0054
 - Specificity:- 0.999
- Test set
 - Accuracy:- 0.965
 - Sensitivity:- 0.0103
 - Specificity:- 0.999

Overall, the model is performing well in the test set, what it had learnt from the train set.

Final conclusion with no PCA

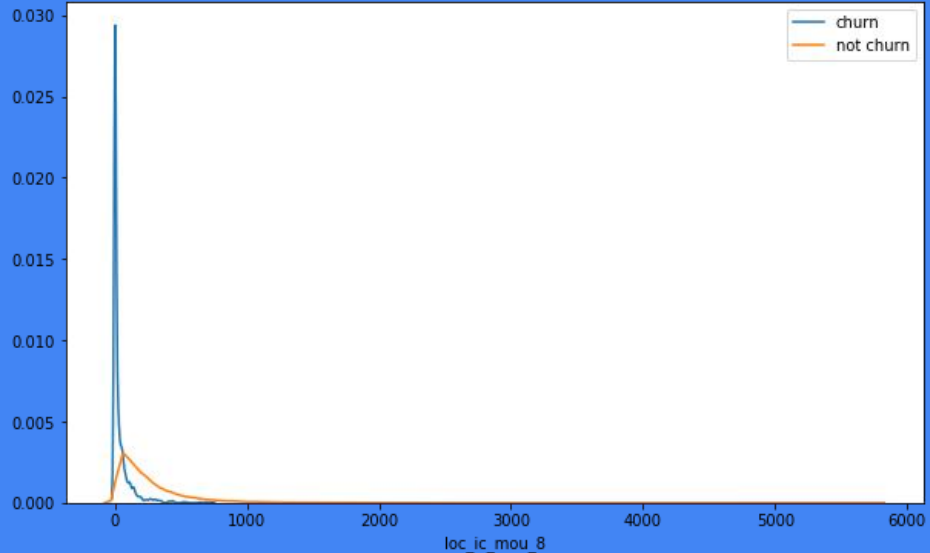
Logistic model with no PCA has good sensitivity and accuracy, as comparison to the models with PCA.

Business Recommendation

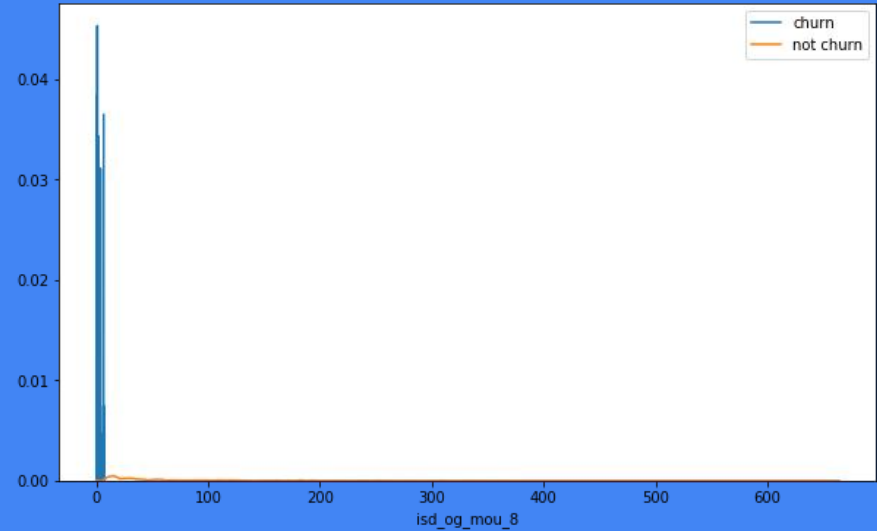
- *Target the customers, whose MOU of the incoming local calls and outgoing ISD calls are less in the action phase.*
- *Target the customers, whose outgoing others charge in July and incoming others on August are less.*
- *customers having value based cost in the action phase increased are more likely to churn than the other customers,so these customers can be a good target to provide offer*
- *Cutomers decreasing monthly 2g usage for August are most probable to churn.*
- *Cutomers, whose monthly 3G recharge in August is more, are likely to be churned.*
- *Customers having decreasing STD incoming minutes of usage for operators of the month of August are more likely to churn.*
- *customers, whose roaming outgoing minutes of usage is increasing are more likely to churn.*
- *Customers having decreasing incoming minutes of usage for operators of the August are more likely to churn.*

Plots of important predictors for churn and non churn customers

We can see that for the beat clients the minutes of use for the period of August is generally populated on the lower side than the non stir clients.

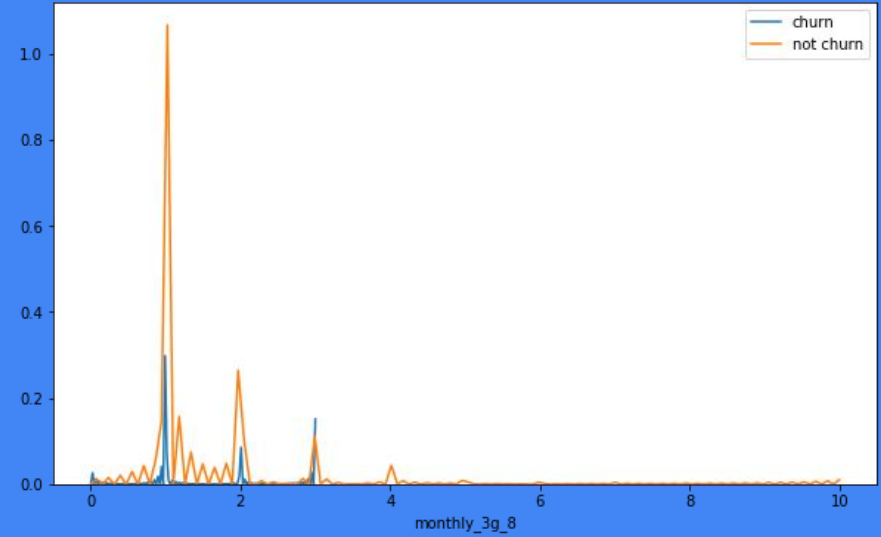


We can see that for the beat clients the minutes of use for the long stretch of August is generally populated on the lower side than the non stir clients.

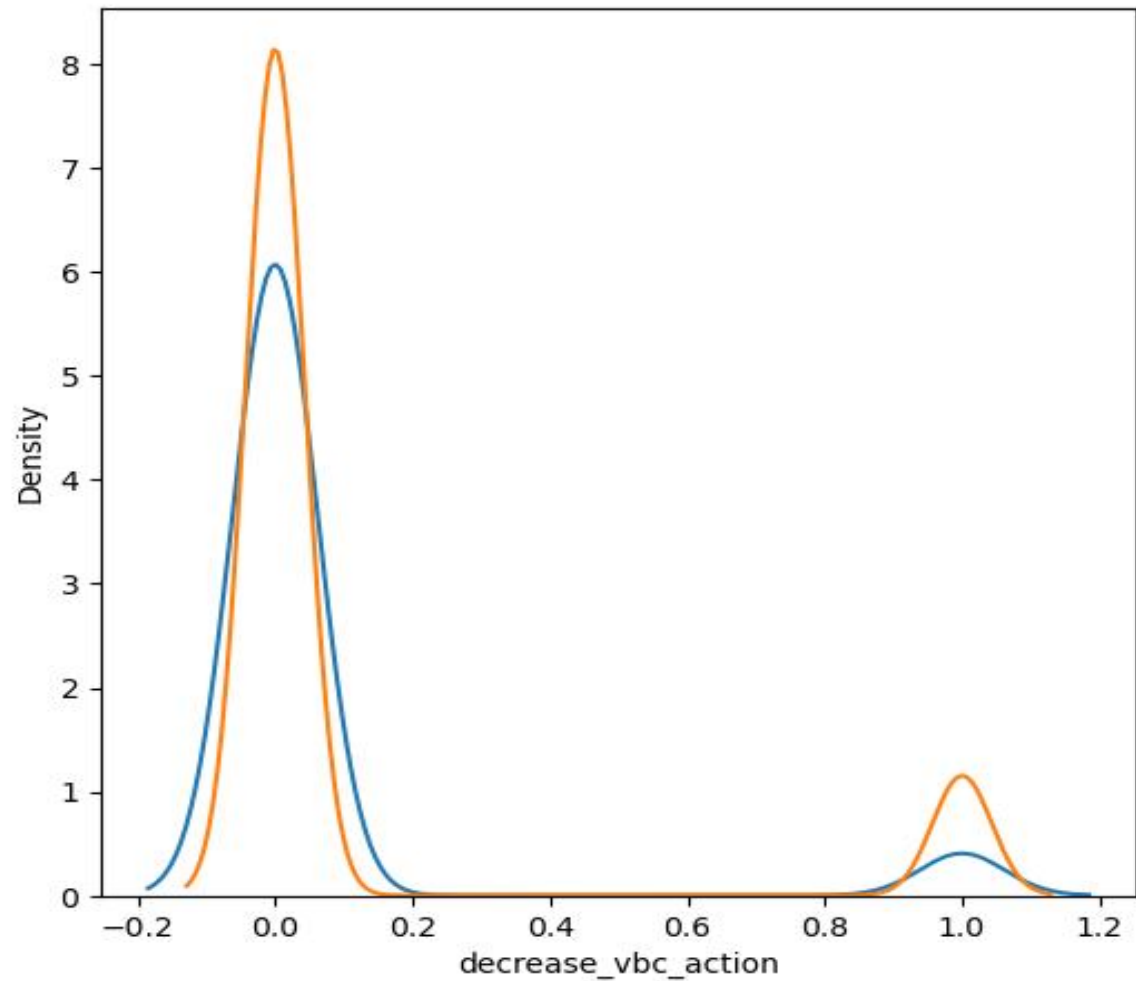


The quantity of monthly 3g information for August for the churn clients are especially populated around 1, though of non churn clients it spreaded across different numbers.

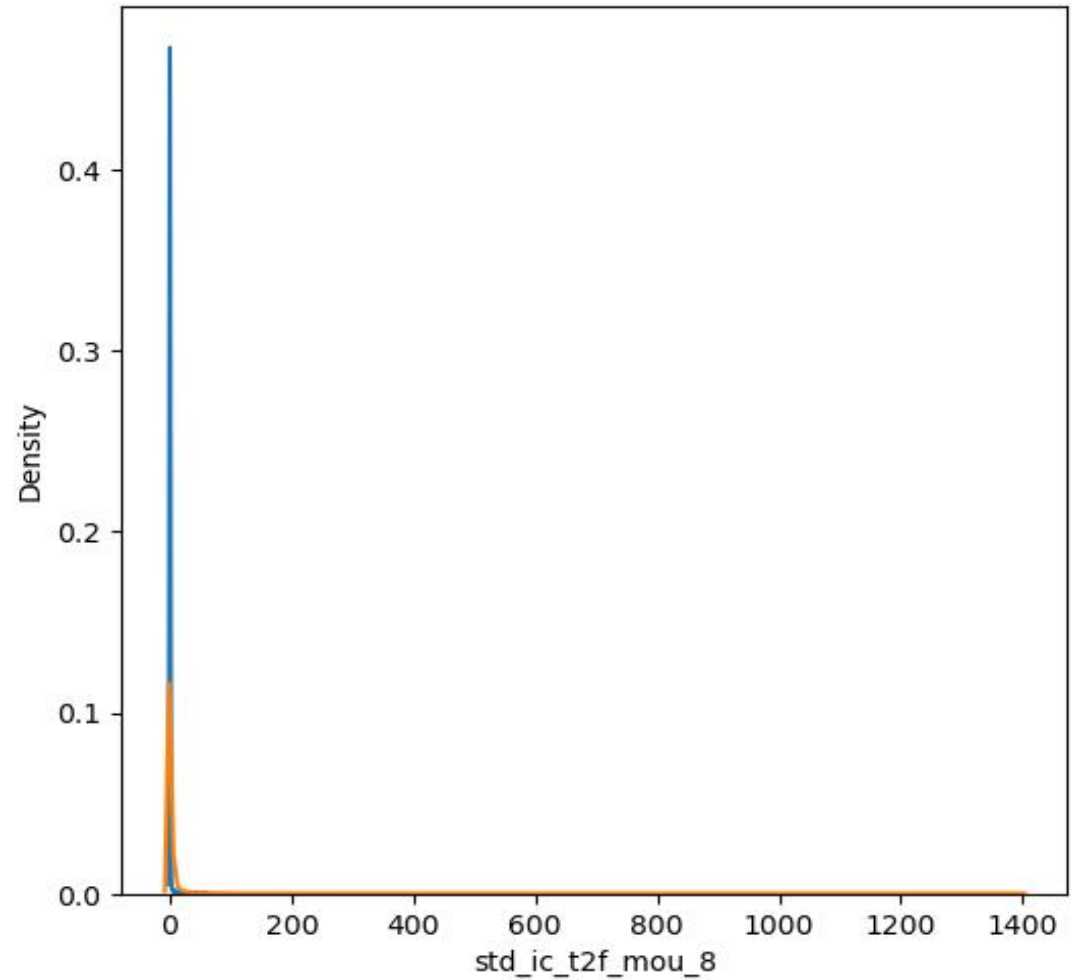
Comparably we can plot every factors, which have higher coefficients, stir appropriation.



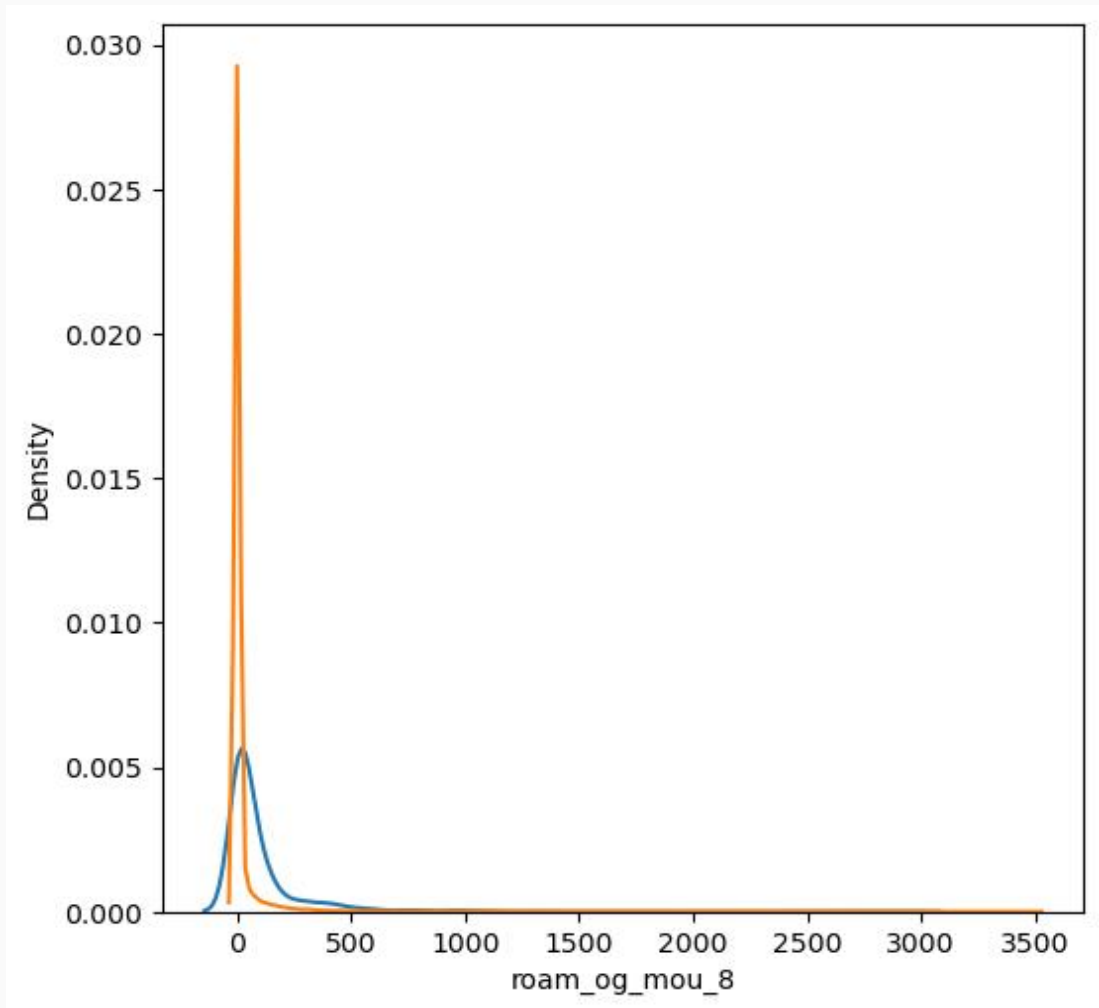
- *Plotting decrease_vbc_action for churn and not churn customers*



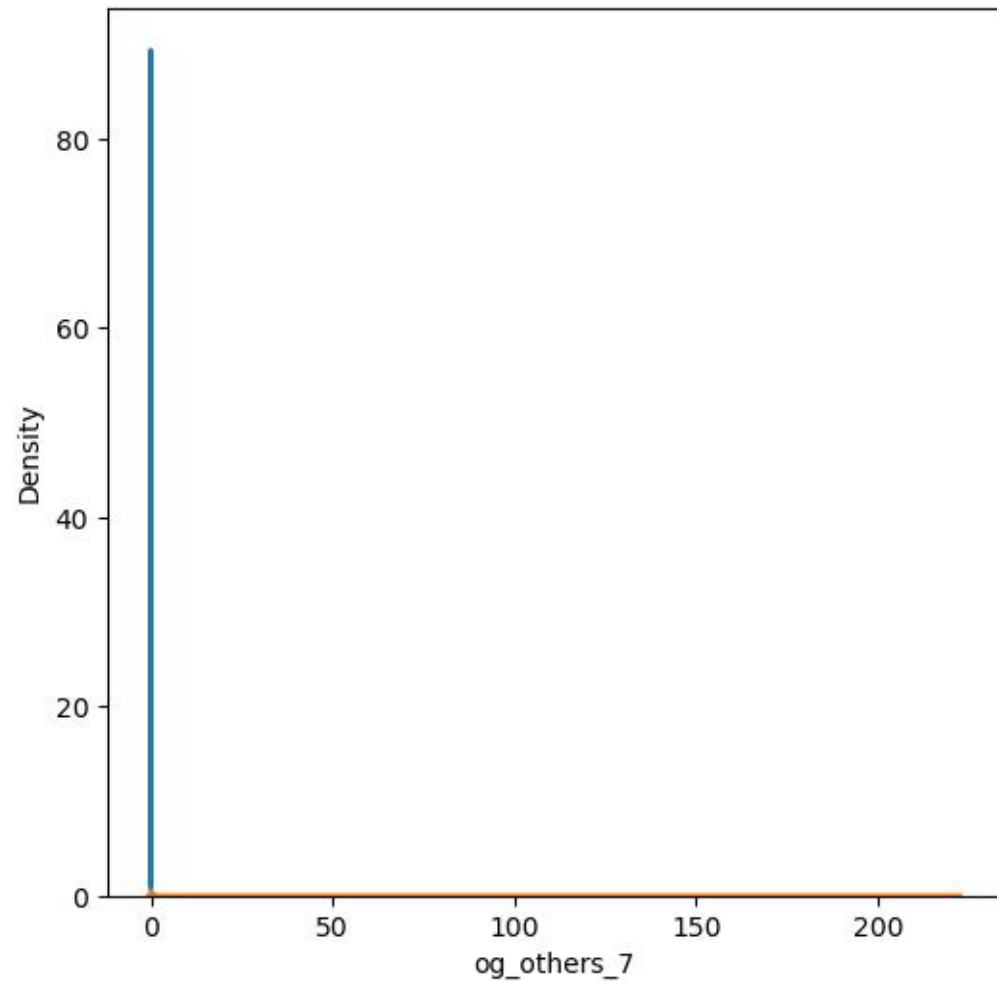
*Plotting std_ic_t2f_mou_8 for churn
and not churn customers*



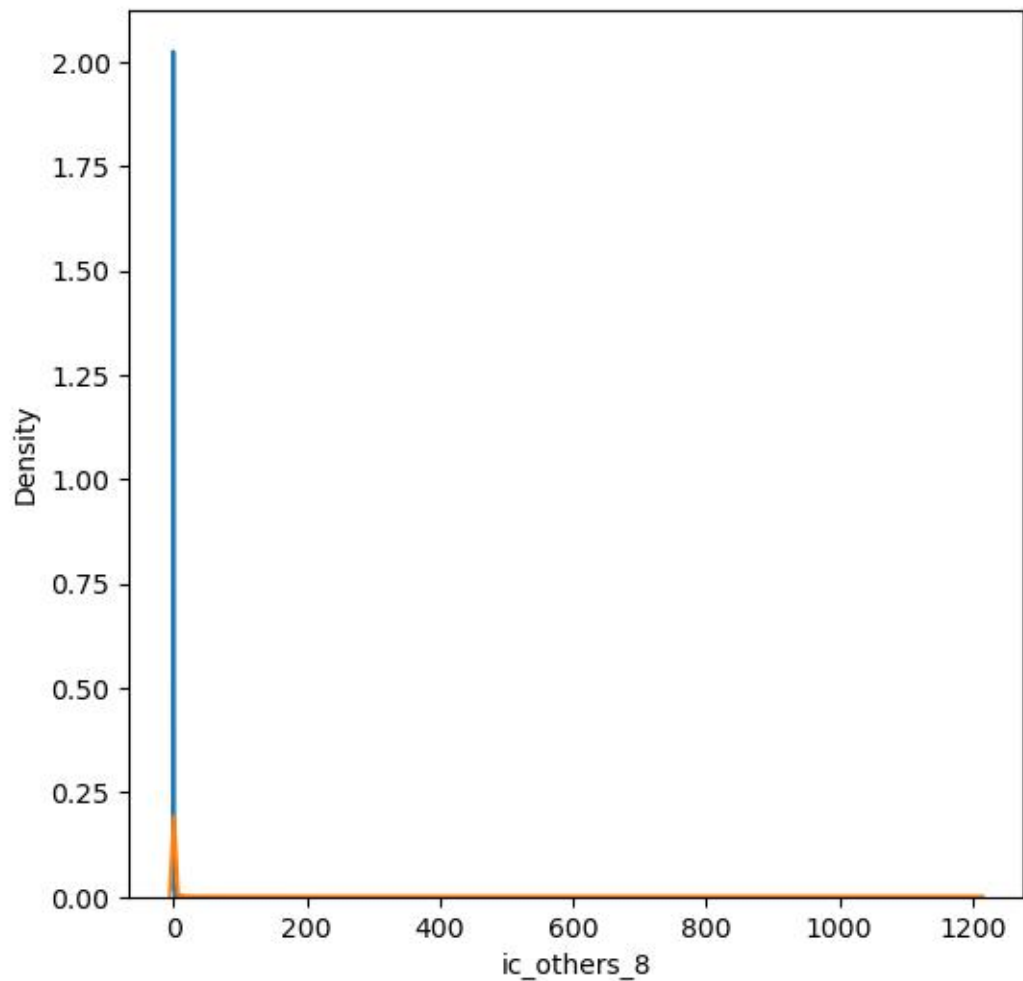
- *Plotting roam_og_mou_8 for churn and not churn customers*



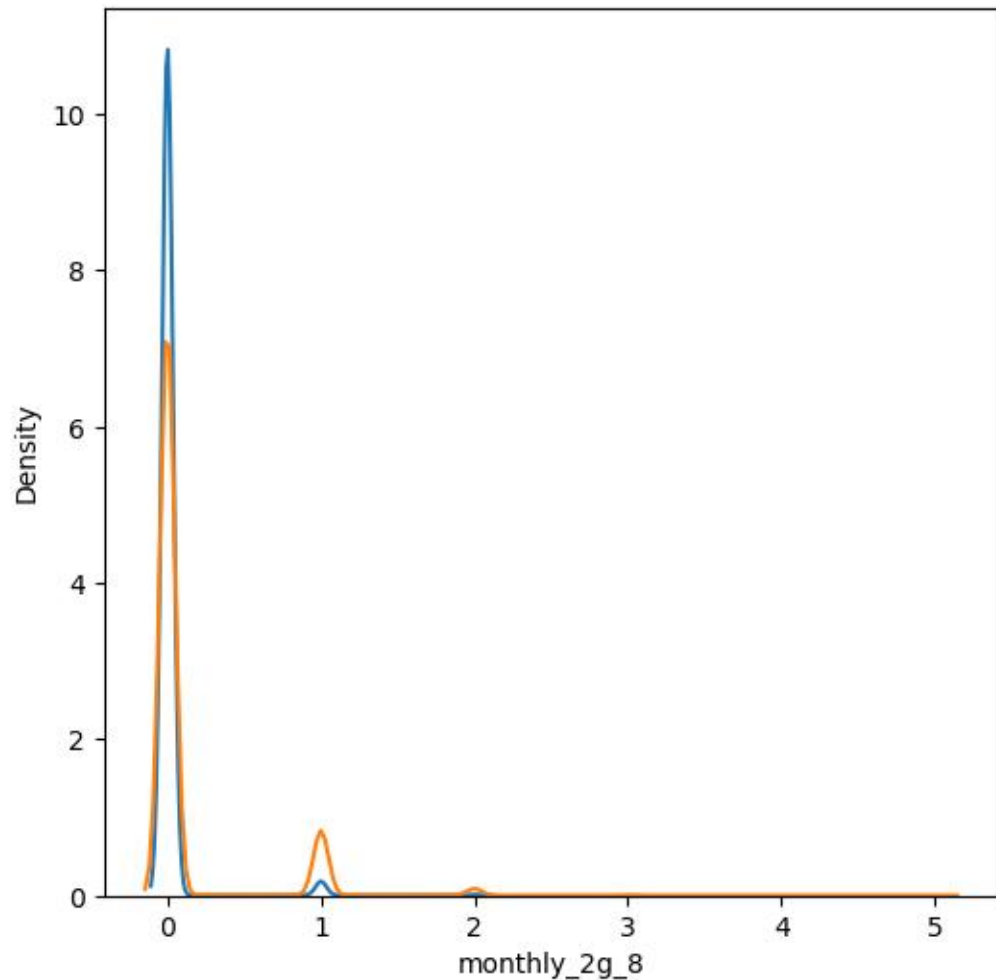
- *Plotting og_others_7 for churn and not churn customers*



- *Plotting ic_others_8 for churn and not churn customers*



- *Plotting monthly_2g_8 for churn and not churn customers*



- *Plotting loc_ic_t2f_mou_8 for churn and not churn customers*

