# *Telecom Churn Case Study*


- Saravanan Ilangovan
- Sathiyaraj S
- Sanjay Katti

## Problem Statement

### 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, we 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

## Data Preparation

1. **Derive new features** This is one of the most important parts of data preparation since good features are often the differentiators between good and bad models. We will use our business understanding to derive features that we think could be important indicators of churn.
2. **Filter high-value customers** As mentioned above, we need to predict churn only for the high-value customers. Define high-value customers as follows: Those who have recharged with an amount more than or equal to X, where X is the 70th percentile of the average recharge amount in the first two months (the good phase).
3. **Tag churners and remove attributes of the churn phase** Now tag the churned customers (churn=1, else 0) based on the fourth month as follows: Those who have not made any calls (either incoming or outgoing) AND have not used mobile internet even once in the churn phase. The attributes we need to use to tag churners are:

    - total_ic_mou_9
    - total_og_mou_9
    - vol_2g_mb_9
    - vol_3g_mb_9 After tagging churners, we need to remove all the attributes corresponding to the churn phase (all attributes having '_9', etc. in their names)

## Objective of the Case study

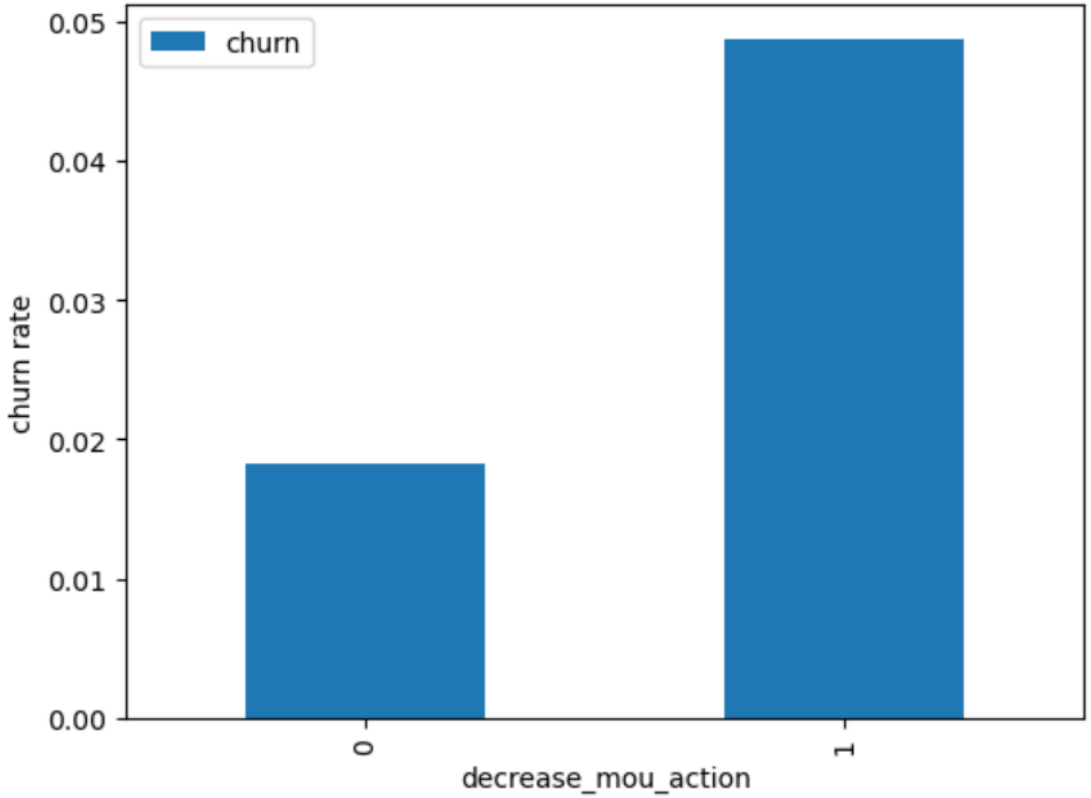
The **business objective** is to predict the churn in the last (i.e. the ninth) month using the data (features) from the first three months.

To do this task well, understanding the typical customer behavior during churn will be helpful.

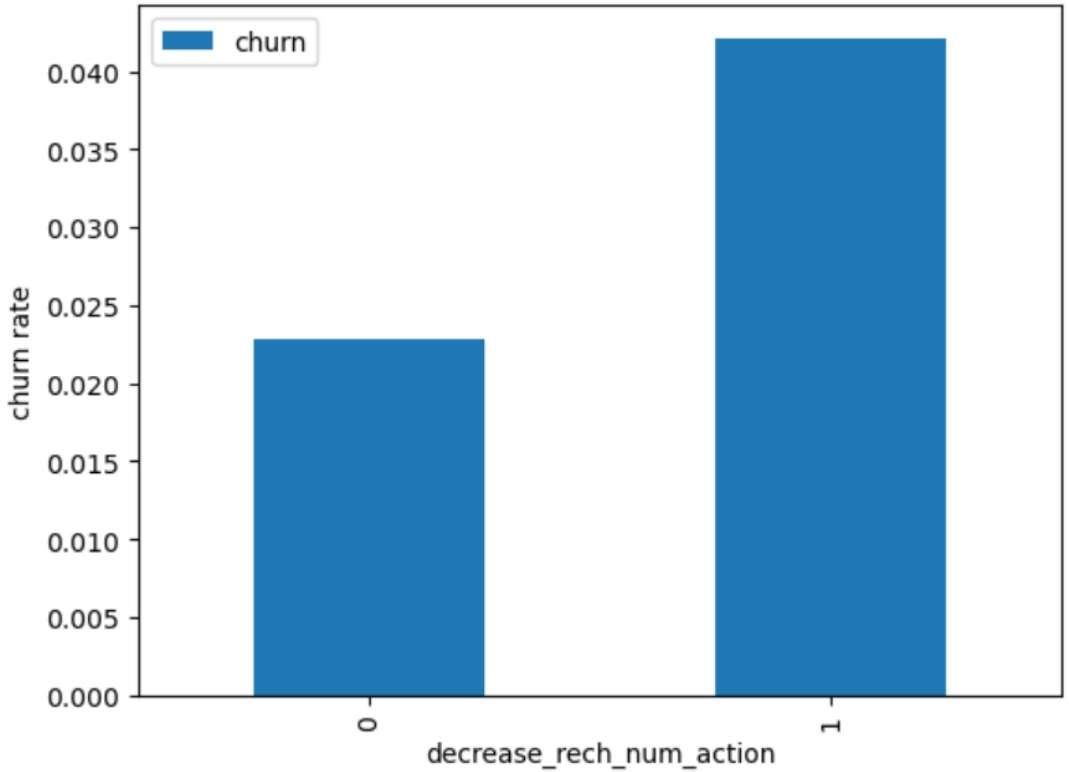
# Approach

- Libraries needed for data cleansing and visualization are imported.
- We dropped columns that have more than 30% null values
- We dropped columns like circle_id, and date columns which are not of importance
- As suggested we filtered the hig,h value customers using the 70% percentile and we found we have more than ~30K records
- We dropped rows and columns having missing values more than 50%
- We analyzed the Null, NA ana d NAN values and removed them, we lost around 7% of records.
- Outliers identified are handled using the IQR method, outliers below the $10^{th}$ percentile and above 90% were removed.
- We derived new features for better understanding of the data
- Data imbalance is analyzed for the Target column and plotted for better understanding.
- Univariate/Bivariate Analysis of the relevant Categorical/numerical is done and insights are derived
- We are more focused on higher Sensitivity/Recall scores than the accuracy.
- We need to care more about churn cases than the not churn cases. The main goal is to retain the customers, who have the possibility to churn. There should not be a problem, if we consider a few not churn customers as churn customers and provide them some incentives for retaining them. Hence, the sensitivity score is more important here.
- We created synthetic samples by doing upsampling using SMOTE(Synthetic Minority Oversampling Technique)
- Features selecting using RFE with 15 columns with and Model Building.

# Univariant Analysis

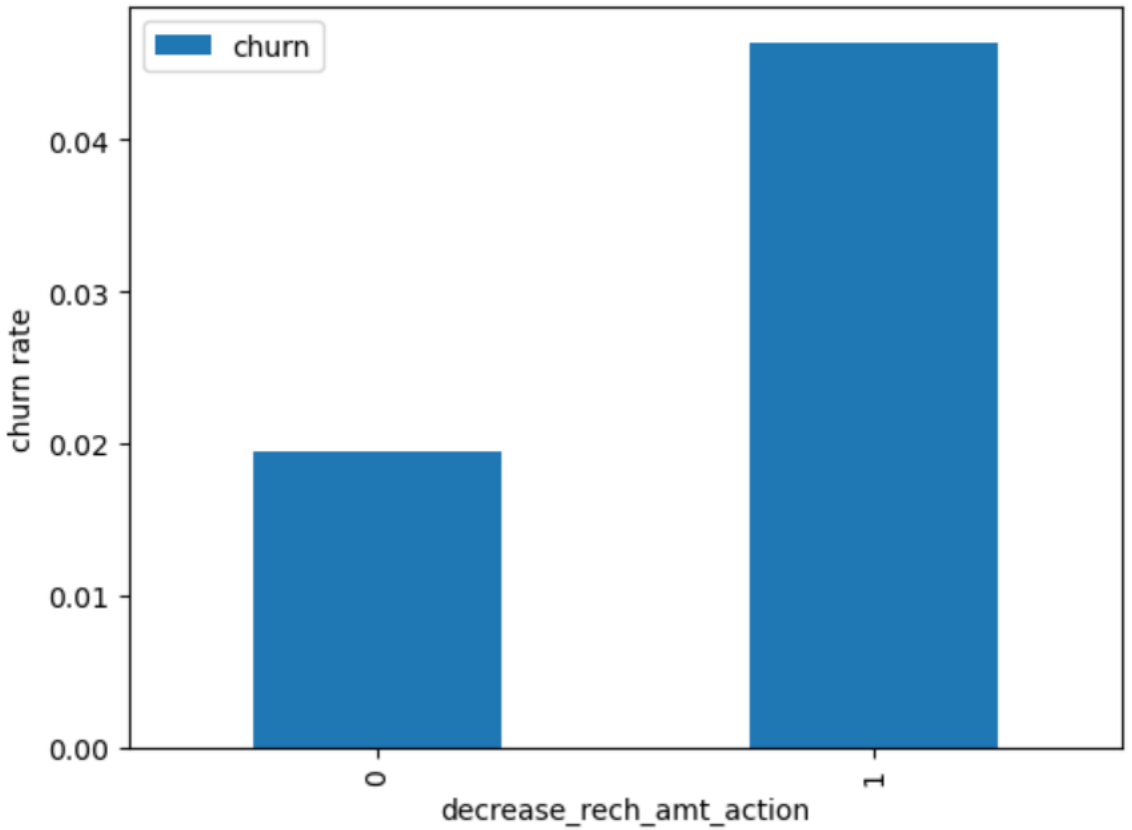



We can see that the churn rate is more for the customers, whose minutes of usage(mou) decreased in the action phase than the good phase.

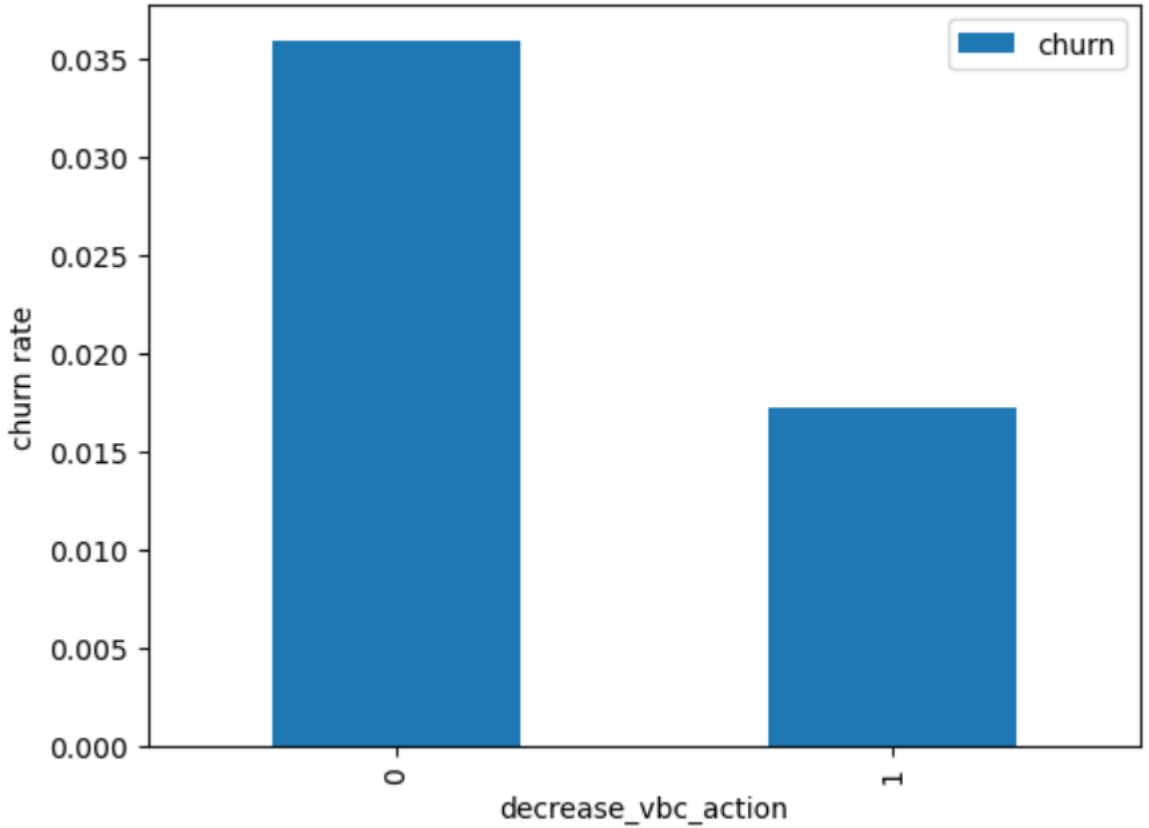



As expected, the churn rate is more for the customers, whose number of recharge in the action phase is lesser than the number in good phase.

# Univariant Analysis

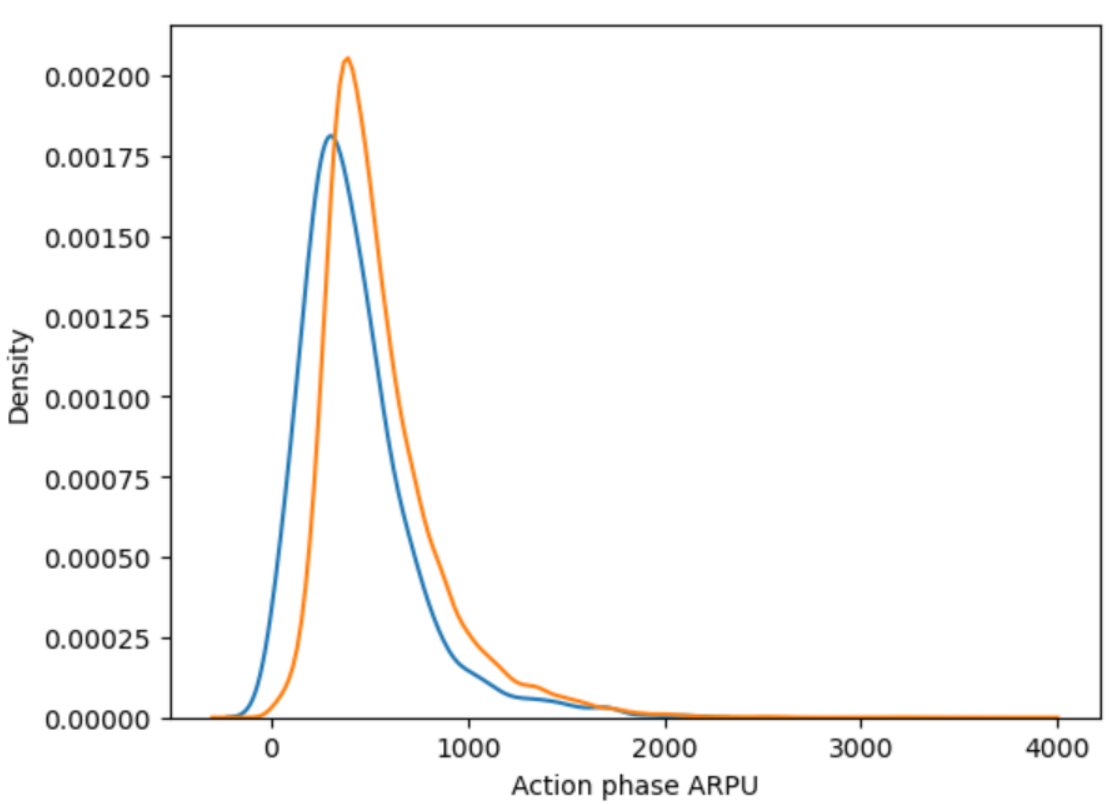



Here also we see the same behaviour. The churn rate is more for the customers, whose amount of recharge in the action phase is lesser than the amount in good phase

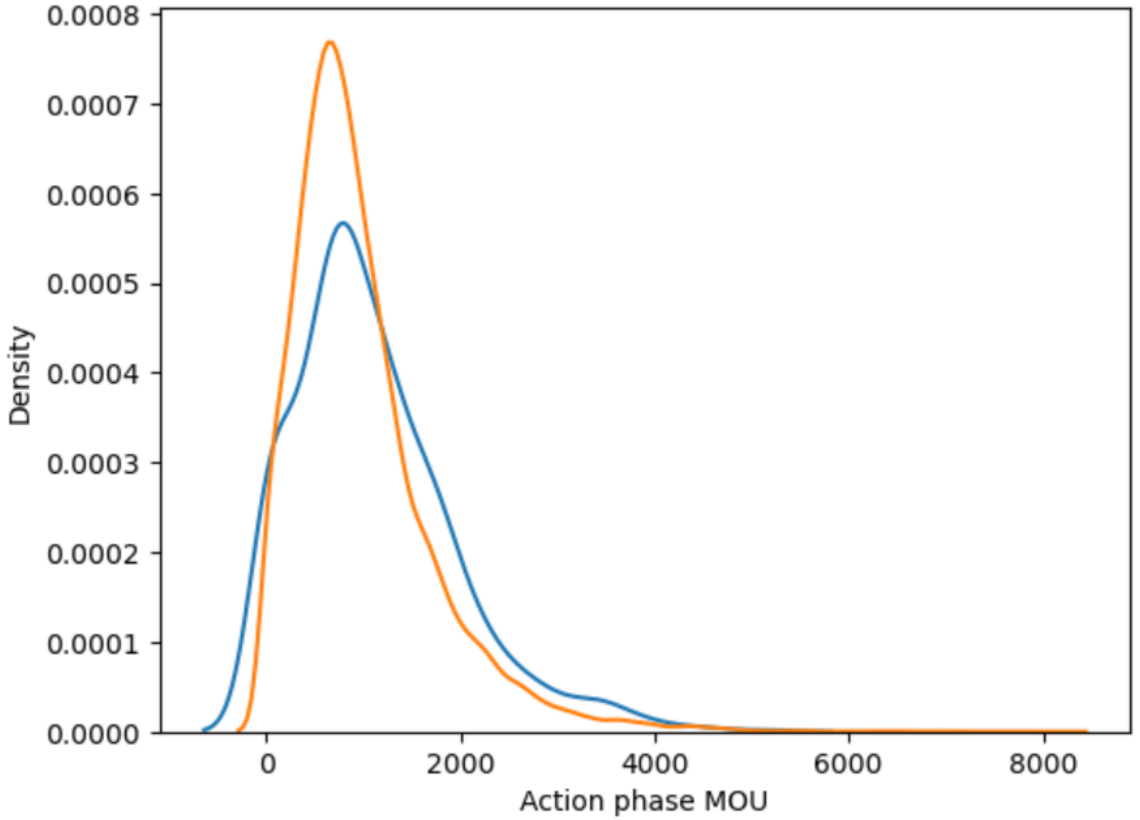



Here we see the expected result. The churn rate is more for the customers, whose volume based cost in action month is increased. That means the customers do not do the monthly recharge more when they are in the action phase.

# Univariant Analysis

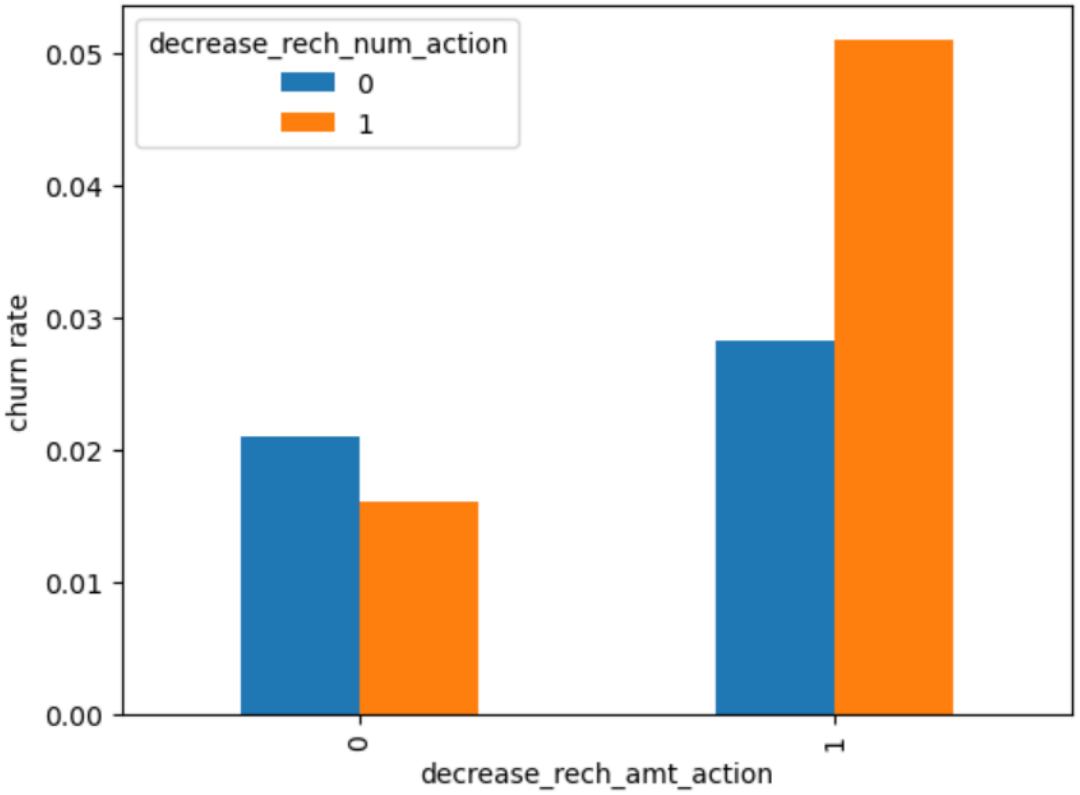


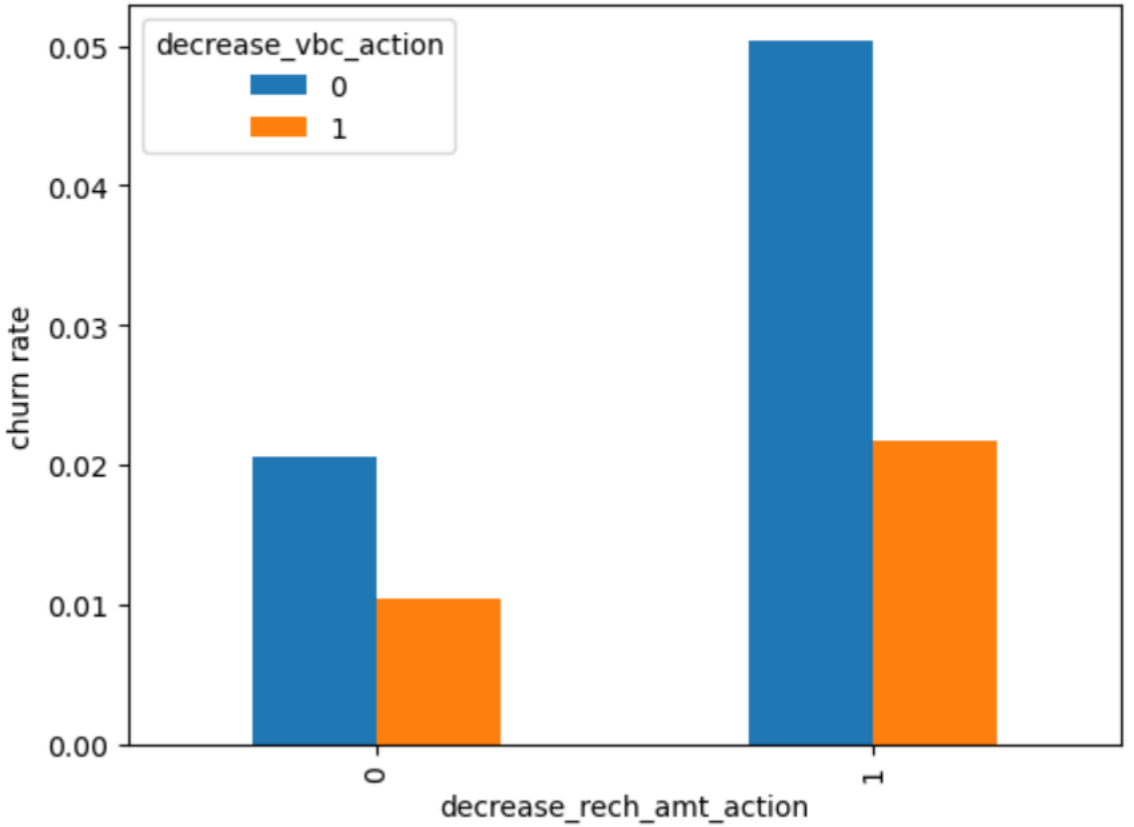

Average revenue per user (ARPU) for the churned customers is mostly densed on the 0 to 900. The higher ARPU customers are less likely to be churned.
ARPU for the not churned customers is mostly densed on the 0 to 1000.




Minutes of usage(MOU) of the churn customers is mostly populated on the 0 to 2500 range. Higher the MOU, lesser the churn probability

# Bivariant Analysis

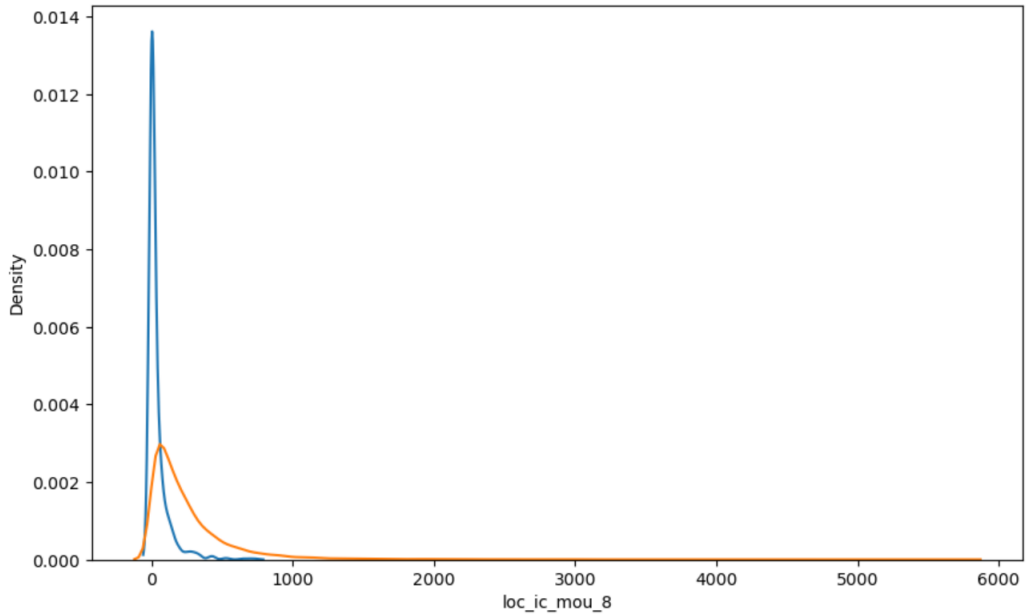


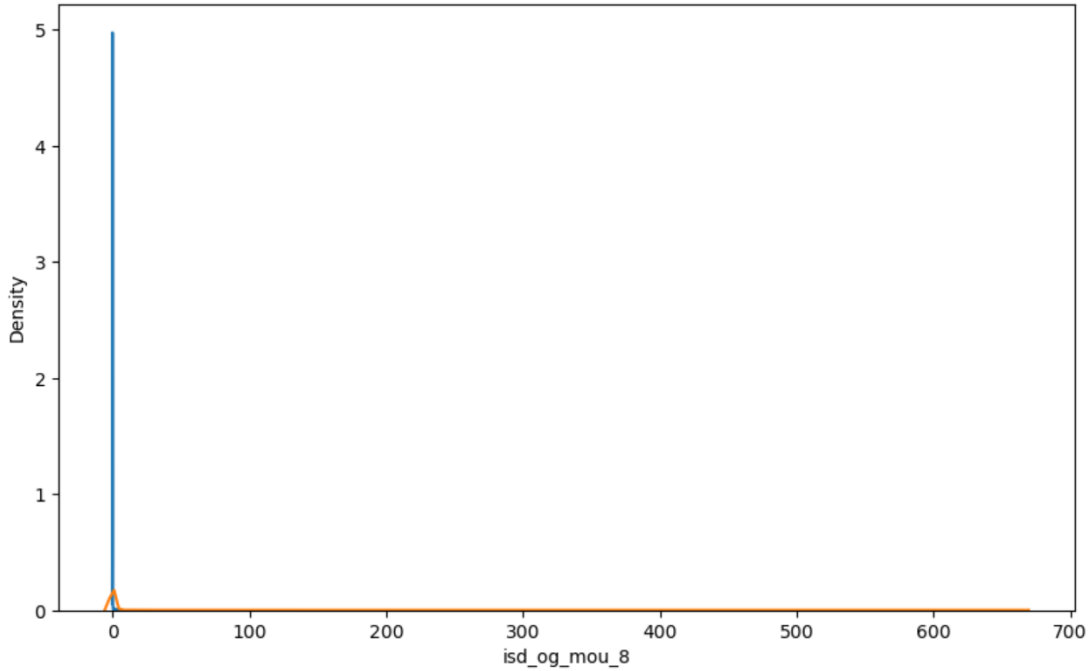

We can see from the above plot, that the churn rate is more for the customers, whose recharge amount as well as number of recharge have decreased in the action phase than the good phase.

we can see that 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.

# Plots of important predictors for churn and non-churn customers

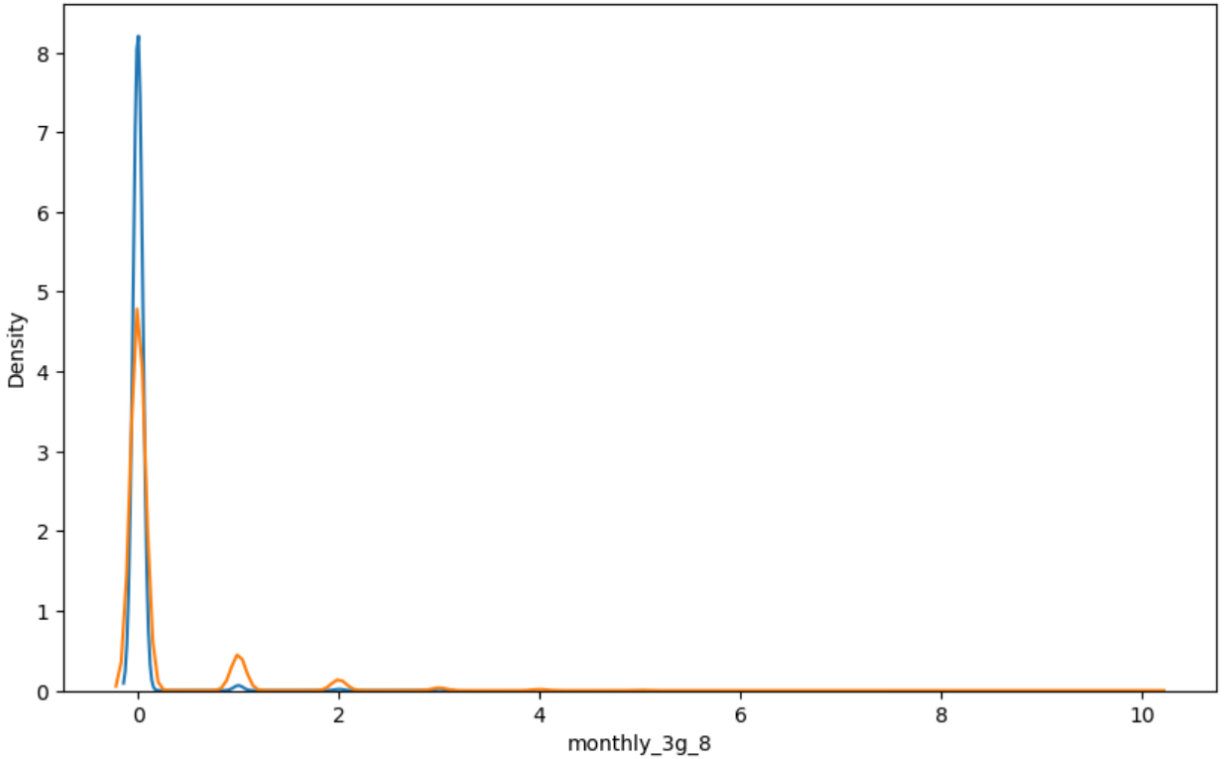



We can see that for the churn customers, the minutes of usage for the month of August are mostly populated on the lower side than the non-churn customers




We can see that the ISD outgoing minutes of usage for the month of August for churn customers is dense approximately zero. On the other hand for the non-churn customers, it is little more than the churn customers.

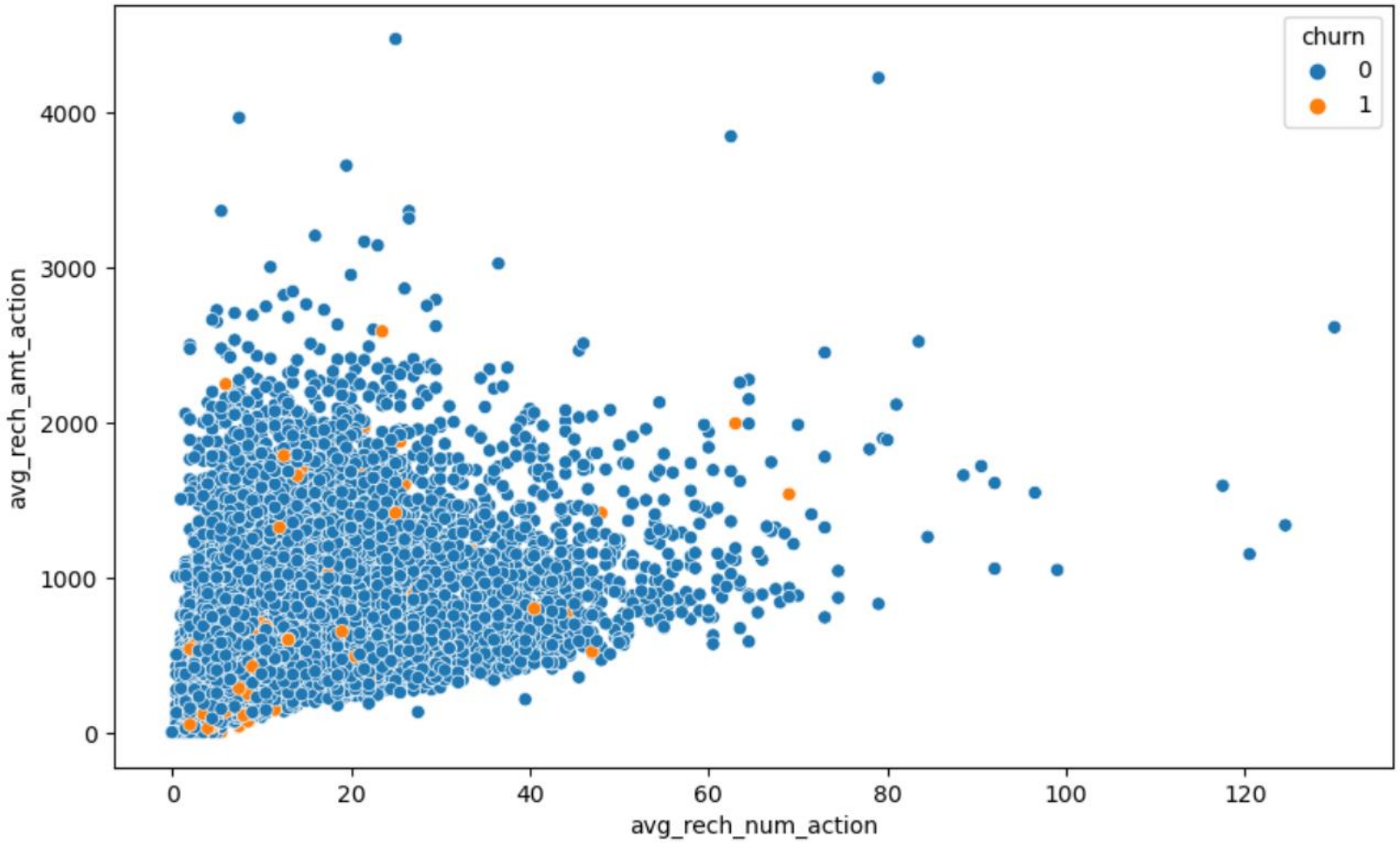
# Plots of important predictors for churn and non-churn customers




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

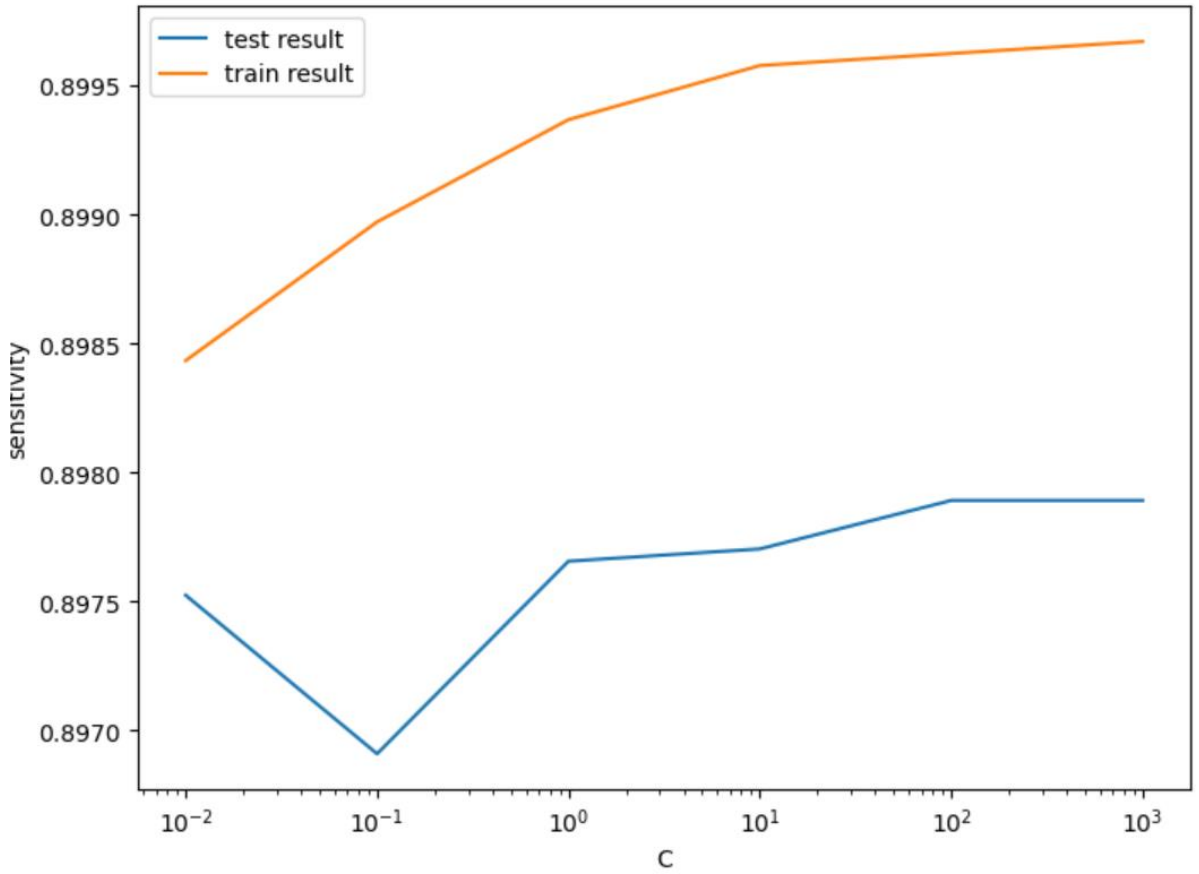
Similarly, we can plot each variable, which has higher coefficients, and churn distribution.

# Bivariant Analysis



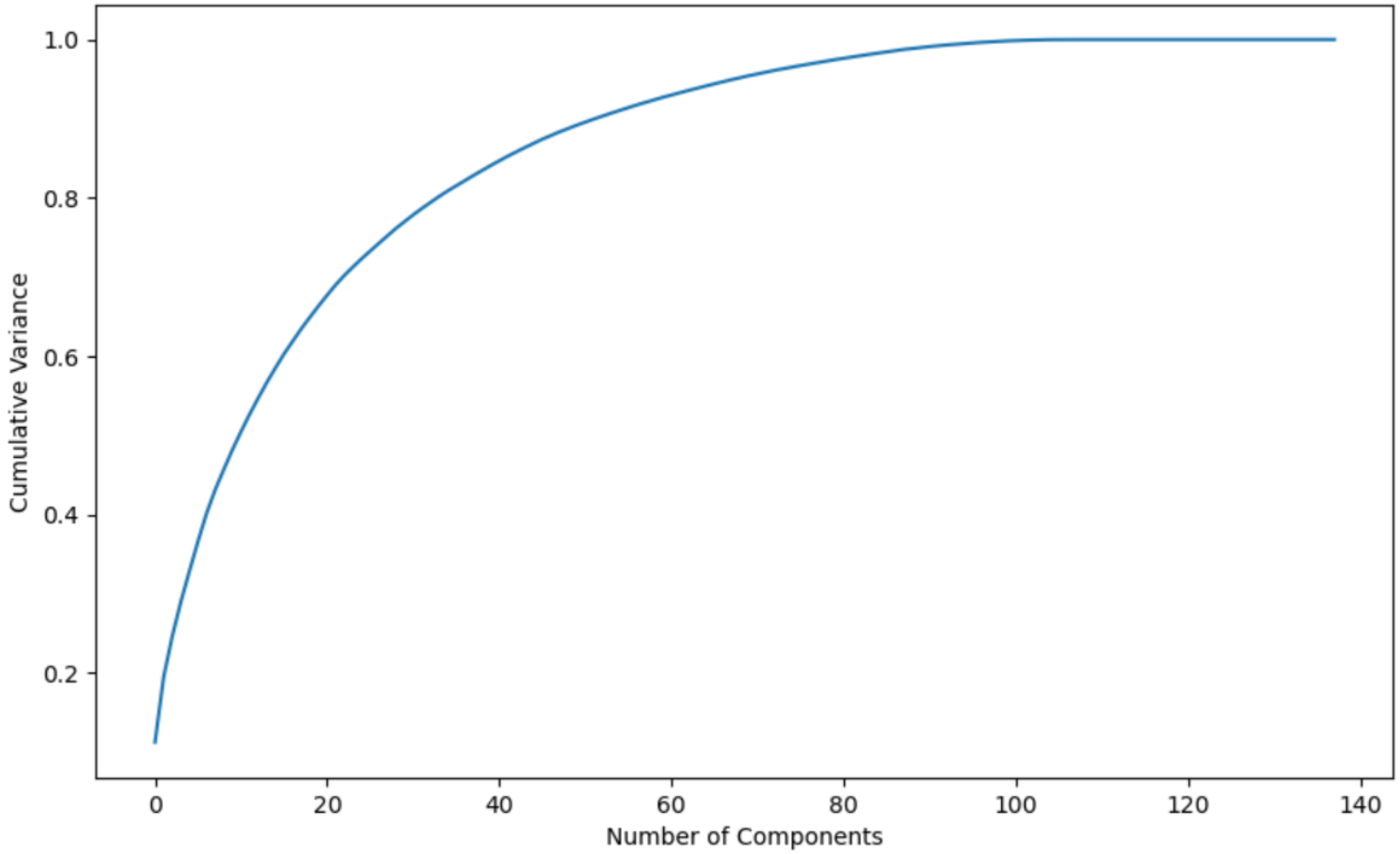

We can see from the above pattern that the recharge number and the recharge amount are mostly proportional. More the number of recharge, more the amount of the recharge

# plot of C versus train and validation scores




C is the inverse of regularization strength in Logistic Regression. Higher values of C correspond to less regularization.

# Bivariant Analysis




We can see that 60 components explain almost more than 90% variance of the data. So, we will perform PCA with 60 components

# Conclusions derived with and without PCA

**Final conclusion with PCA**

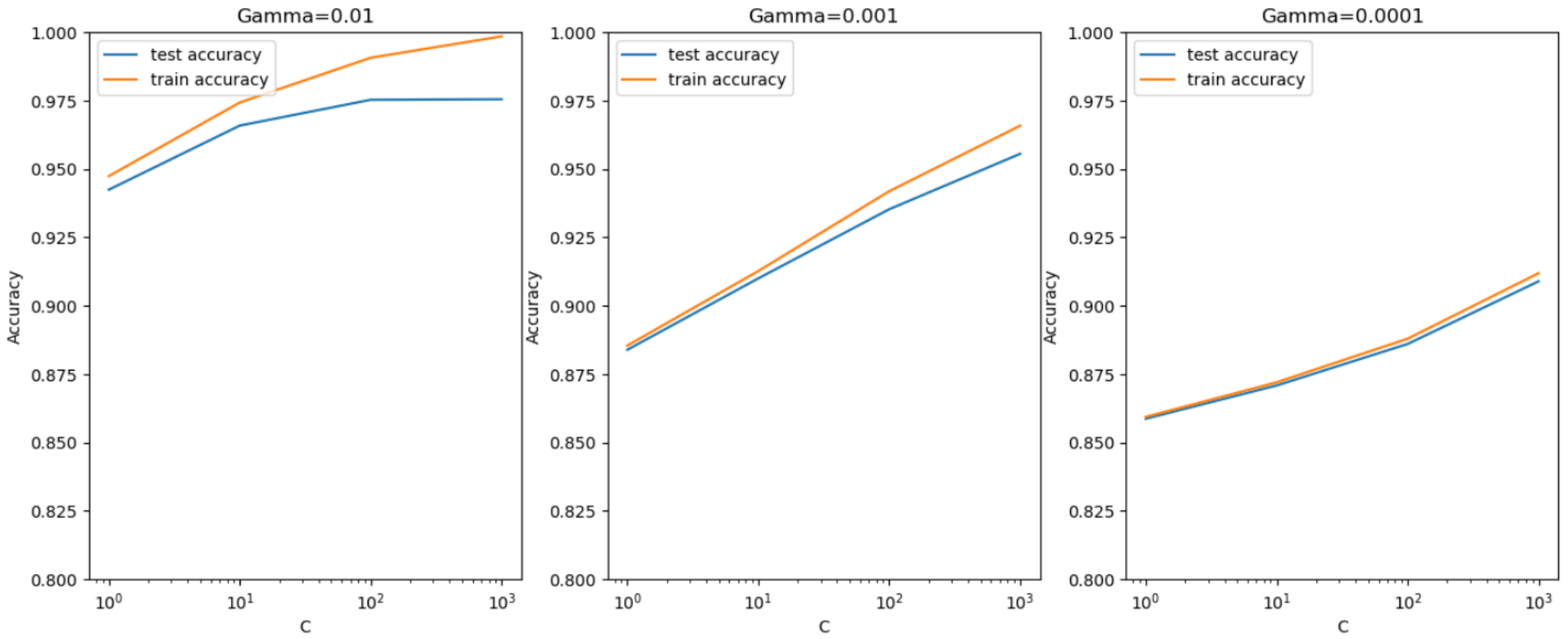
After trying several models we can see that for achieving the best sensitivity, which was our ultimate goal, the classic Logistic regression or the SVM models performs well. For both the models the sensitivity was approx. 81%. Also, we have a good accuracy of approx 85%.

**Final conclusion with no PCA**

We can see that the logistic model with no PCA has good sensitivity and accuracy, which are comparable to the models with PCA. So, we can go for the more simplistic model such as logistic regression with PCA as it explains the important predictor variables as well as the significance of each variable. The model also helps us to identify the variables that should be acted upon for making the decision of the churned customers.

Hence, the model is more relevant in terms of explaining to the business.

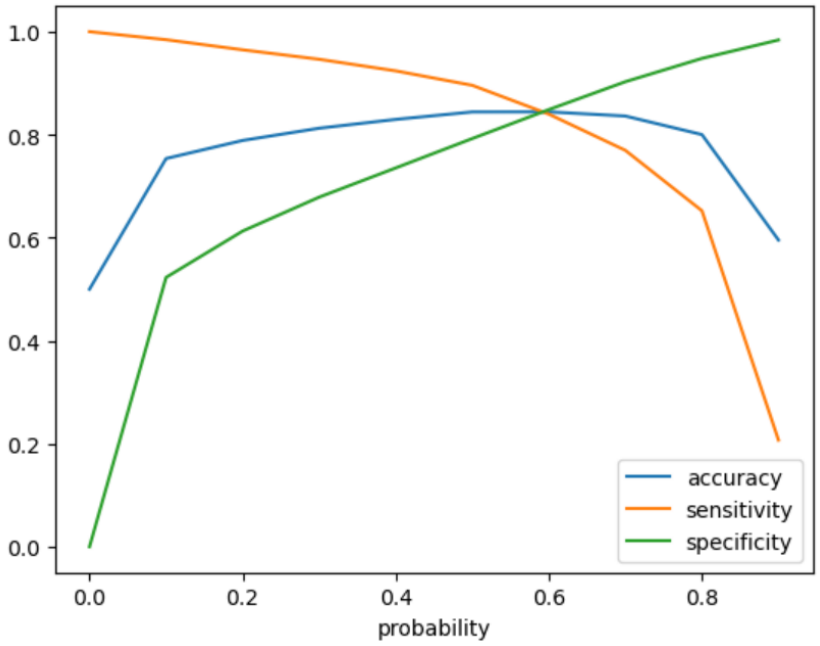
# Plotting the accuracy with various C and gamma values




The best test score is 0.975495911159373 corresponding to hyperparameters {'C': 1000, 'gamma': 0.01}

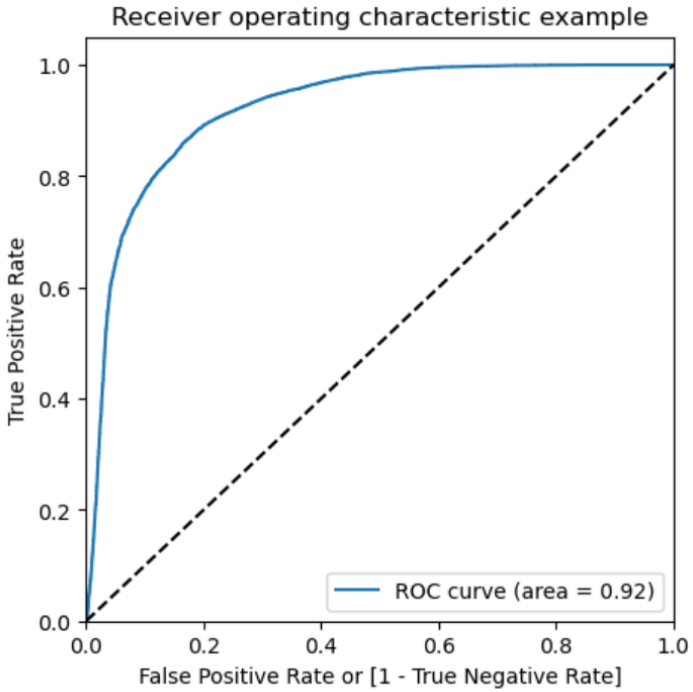
We argue that the model will be simpler if it has as less non-linearity as possible, so we choose gamma=0.0001 and a high C=100

# Accuracy sensitivity and specificity




At point 0.6 where the three parameters cut each other, we can see that there is a balance between sensitivity and specificity with a good accuracy.

# ROC Curve (Trade-off between sensitivity & specificity)




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

# Recommendations

- Top predictors Below are few top variables selected in the logistic regression model.

| Variable | Coefficient |
|---|---|
| loc_ic_mou_8 | -3.3287 |
| og_others_7 | -2.4711 |
| ic_others_8 | -1.5131 |
| isd_og_mou_8 | -1.3811 |
| decrease_vbc_action | -1.3293 |
| monthly_3g_8 | -1.0943 |
| std_ic_t2f_mou_8 | -0.9503 |
| monthly_2g_8 | -0.9279 |
| loc_ic_t2f_mou_8 | -0.7102 |
| roam_og_mou_8 | 0.7135 |

We can see most of the top variables have negative coefficients. That means the variables are inversely correlated with the churn probability.

E.g.:-

If the local incoming minutes of usage (loc_ic_mou_8) are lesser in the month of August than in any other month, then there is a higher chance that the customer is likely to churn.

# Recommendations

- Target the customers, whose minutes of usage of the incoming local calls and outgoing ISD calls are less in the action phase (mostly in the month of August).
- Target the customers, whose outgoing others charge in July and incoming others on August are less.
- Also, the customers having value-based cost in the action phase increased are more likely to churn than the other customers. Hence, these customers may be a good target to provide offer.
- Customers, who's monthly 3G recharge in August is more, are likely to be churned.
- Customers having decreasing STD incoming minutes of usage for operators T to fixed lines of T for the month of August are more likely to churn.
- Customers decreasing monthly 2g usage for August are most probable to churn.
- Customers having decreasing incoming minutes of usage for operators T to fixed lines of T for August are more likely to churn.
- roam_og_mou_8 variables have positive coefficients (0.7135). That means for the customers, whose roaming outgoing minutes of usage is increasing are more likely to churn.