




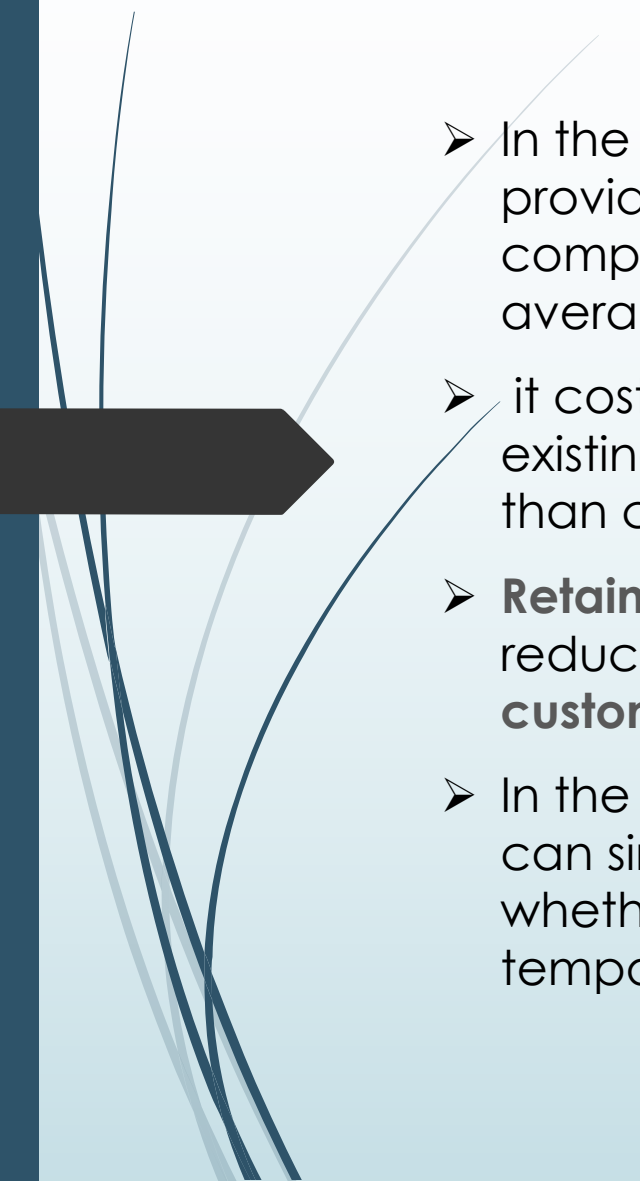
Telecom Churn Case Study- Advanced Machine Learning

By Pradnya Ramesh Gangurde

Content

- 
- **Business Problem Understanding**
 - **Objective**
 - **Solution Methodology**
 - **EDA**
 - **Model building**
 - **Recommendations**
 - **Conclusion**

Business Problem Understanding

- 
- 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.
 - 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.
 - **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 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 which is the most common model in India and Southeast Asia.

Objective

- 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.
- The dataset contains customer-level information for a span of four consecutive months - June, July, August and September. The months are encoded as 6, 7, 8 and 9, respectively. 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.
- Since you are working over a four-month window, the first two months are the 'good' phase, the third month is the 'action' phase, and the fourth month is the 'churn' phase.
- For **Retaining high profitable customers**, we need to analyze 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.

Solution Methodology

➤ **Data cleaning and data manipulation.**

1. Check and handle duplicate data.
2. Check and handle NA values and missing values.
3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
4. Imputation of the values, if necessary.
5. Check and handle outliers in data.

➤ **EDA**

- 1.1. Univariate data analysis: value count, distribution of variable etc.
- 2.2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

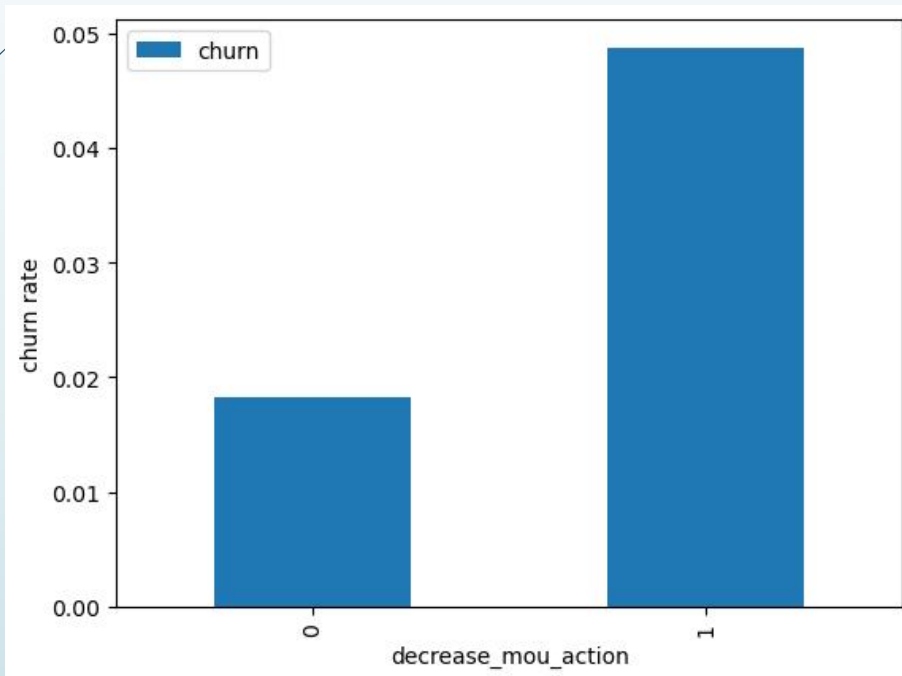
- **Dealing with data imbalance, Feature Scaling & Dummy Variables, encoding of the data.**
- **Classification technique: logistic regression, Decision Tree and Random forest are used for the model making and prediction.**
- **Validation of the model.**
- **Model presentation.**
- **Conclusions and recommendations.**

EDA

A) Univariate analysis

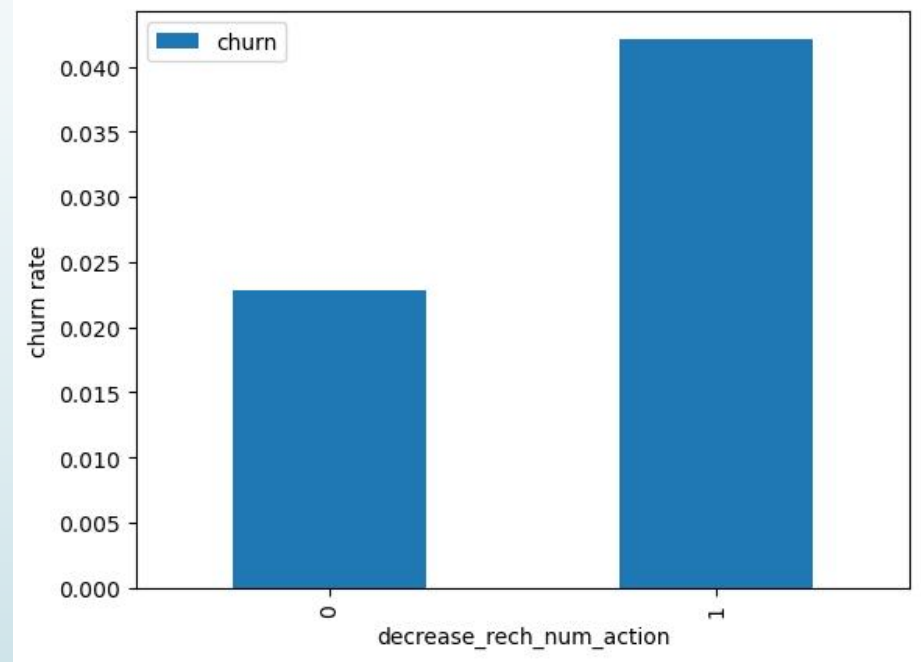
Churn rate on the basis whether the customer decreased her/his MOU in action month

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.



Churn rate on the basis whether the customer decreased her/his number of recharge in action month

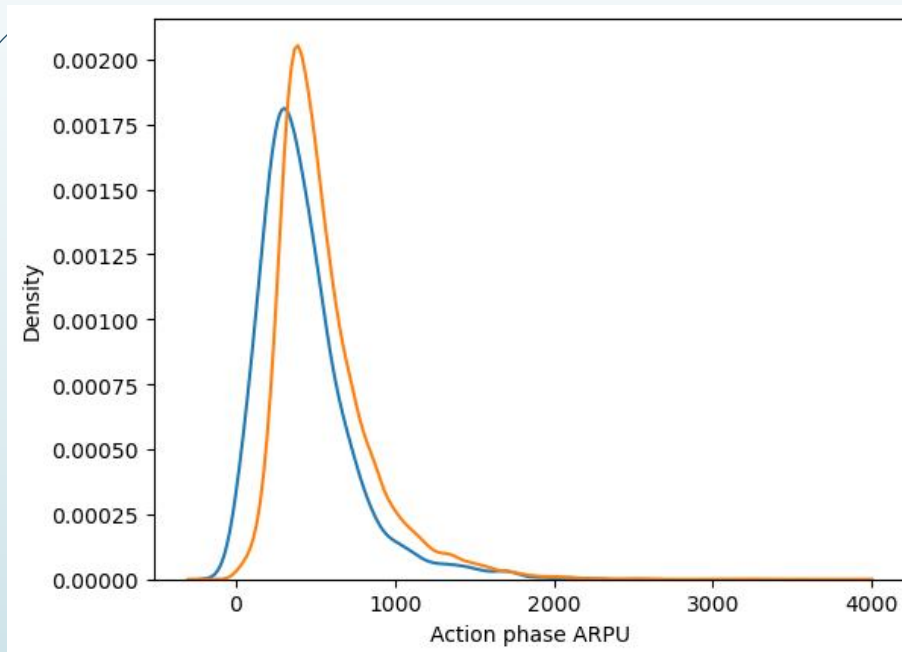
As expected, the churn rate is more for the customers, whose number of recharge in the action phase is lesser than the number in good phase.



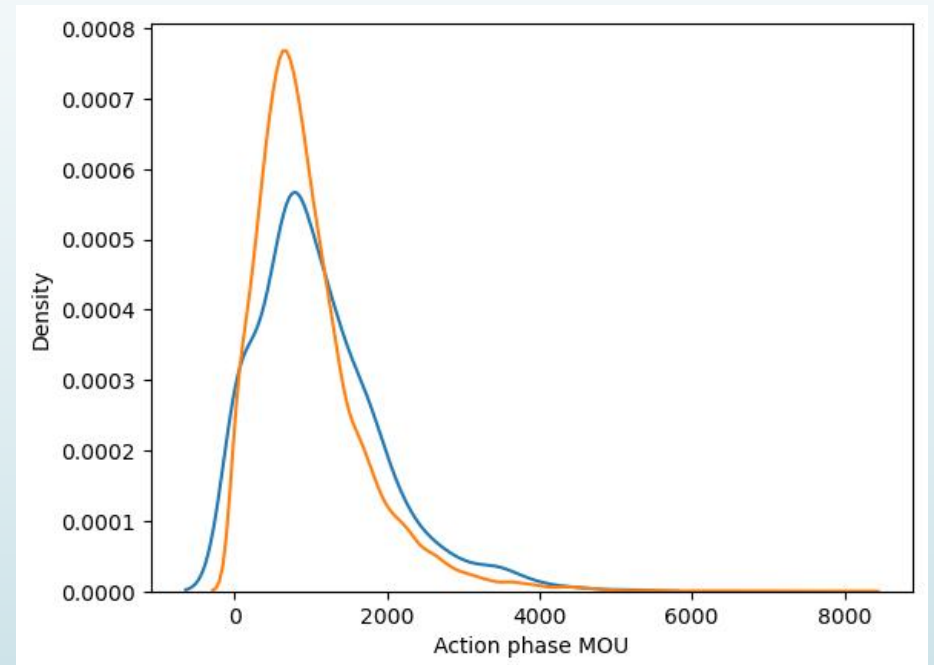
Analysis of the average revenue per customer (churn and not churn) in the action phase :

Below, Average revenue per user (ARPU) for the churned customers is mostly dense on the 0 to 900. The higher ARPU customers are less likely to be churned.

ARPU for the not churned customers is mostly dense on the 0 to 1000.

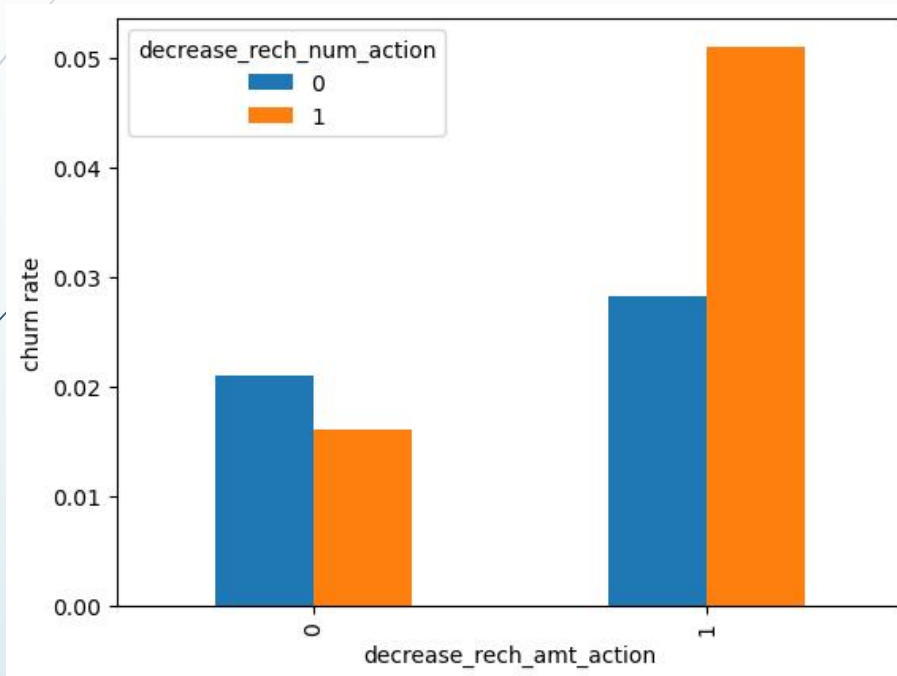


Below, Minutes of usage(MOU) of the churn customers is mostly populated on the 0 to 2500 range. Higher the MOU, lesser the churn probability.

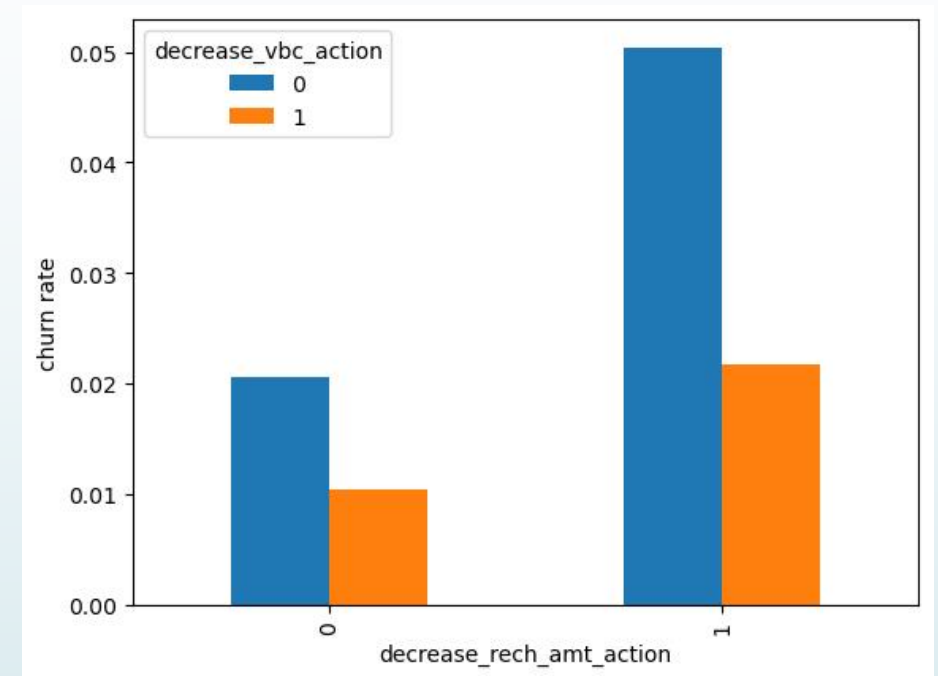


B) Bivariate analysis:

Analysis of churn rate by the decreasing recharge amount and number of recharge in the action phase



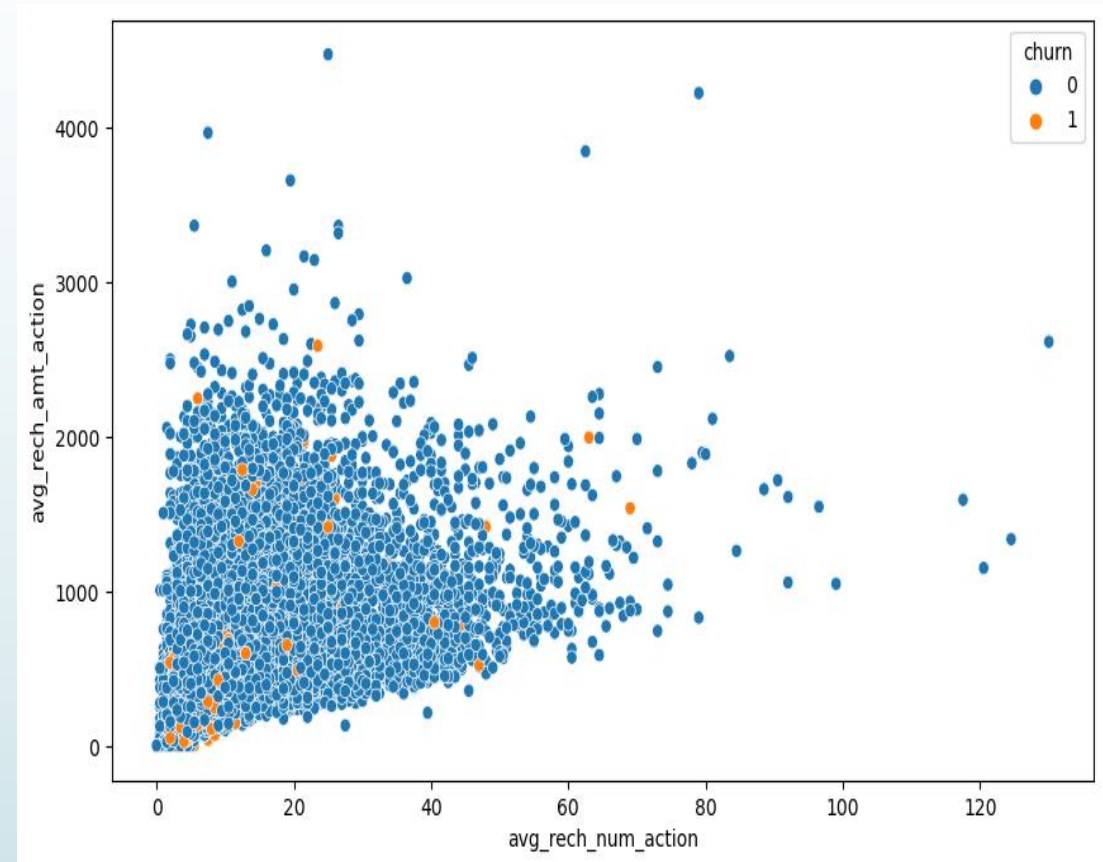
We can see from the above plot, that the churn rate is more for the customers, whose recharge amount as well as number of recharge have decreased in the action phase than the good phase.



Here, also 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.

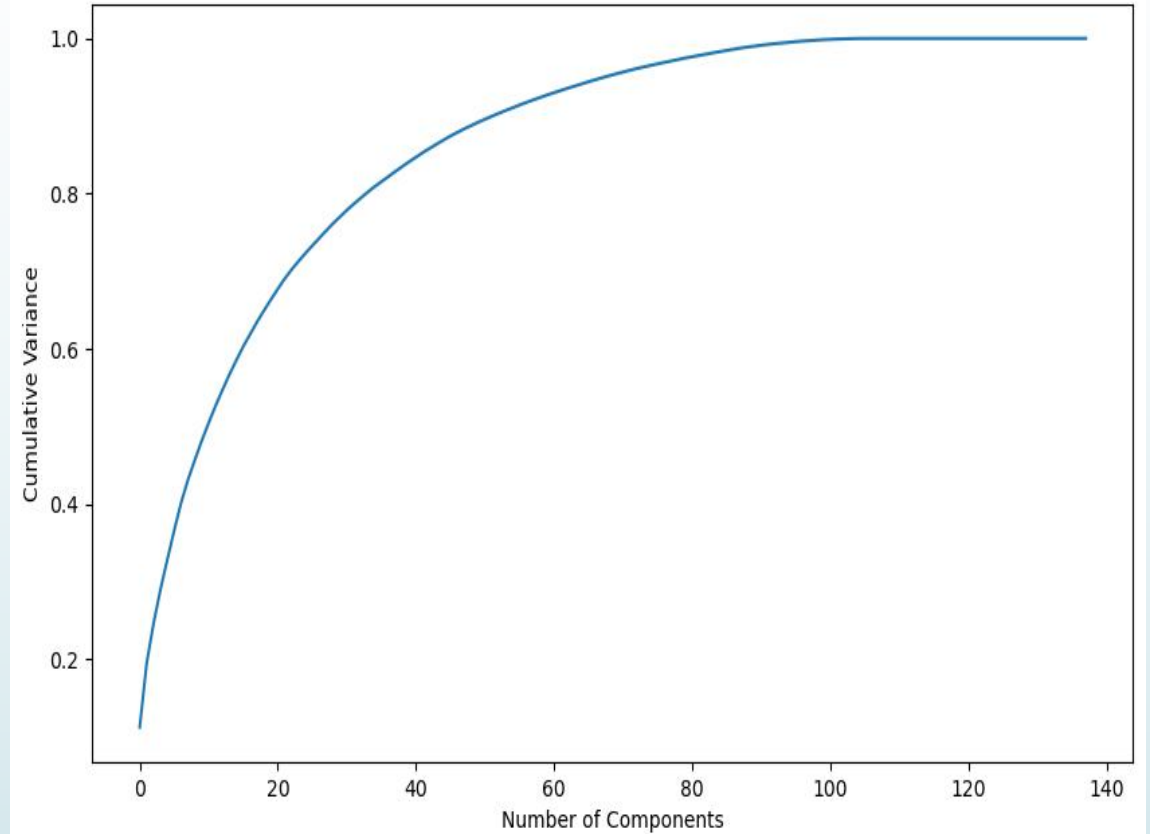
Analysis of recharge amount and number of recharge in action month

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



Model with PCA

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

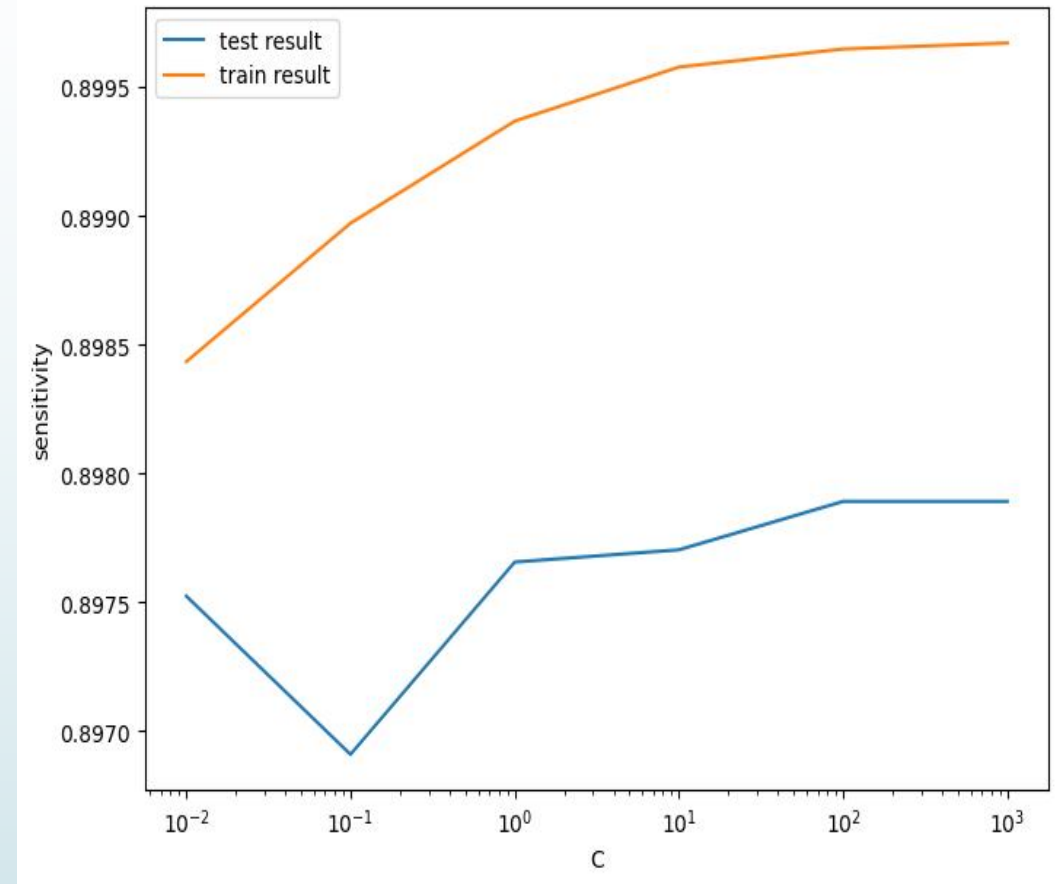


Logistic regression with PCA

Tuning hyperparameter C

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

The highest Test sensitivity is 0.89 at $C = 100$.



The Accuracy, Sensitivity and Specificity for various probability cutoffs.

Analysis of the curve

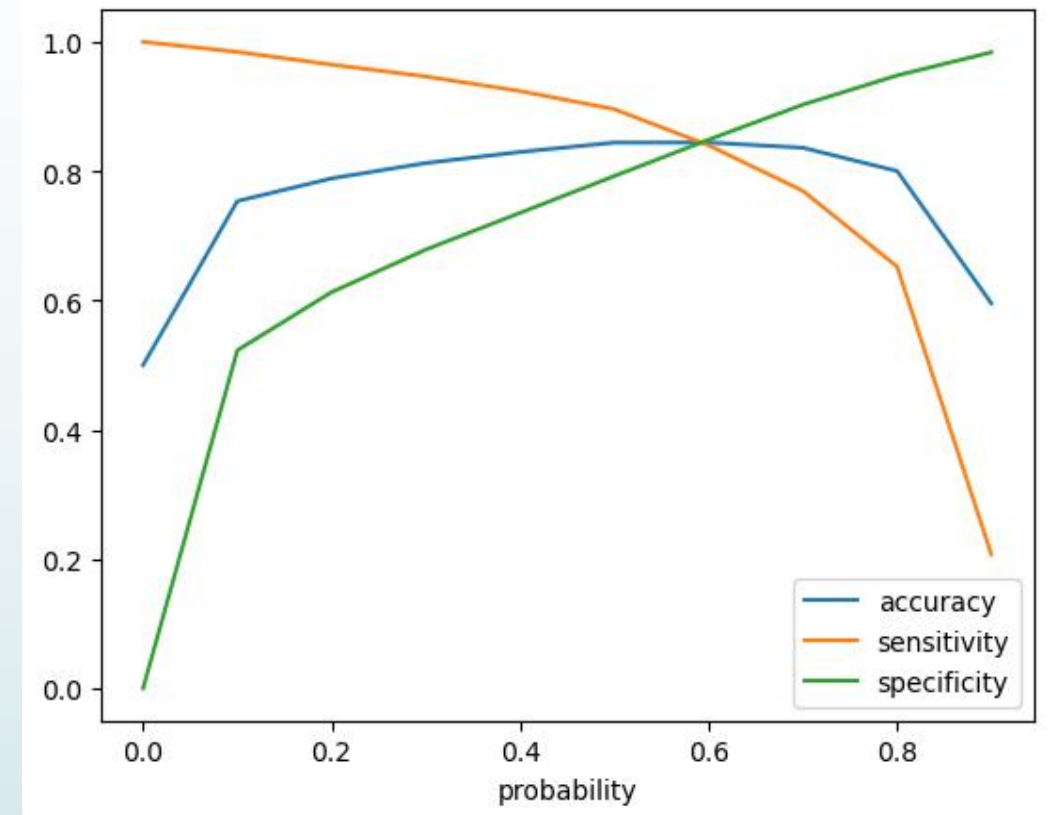
Accuracy - Becomes stable around 0.6

Sensitivity - Decreases with the increased probability.

Specificity - Increases with the increasing probability.

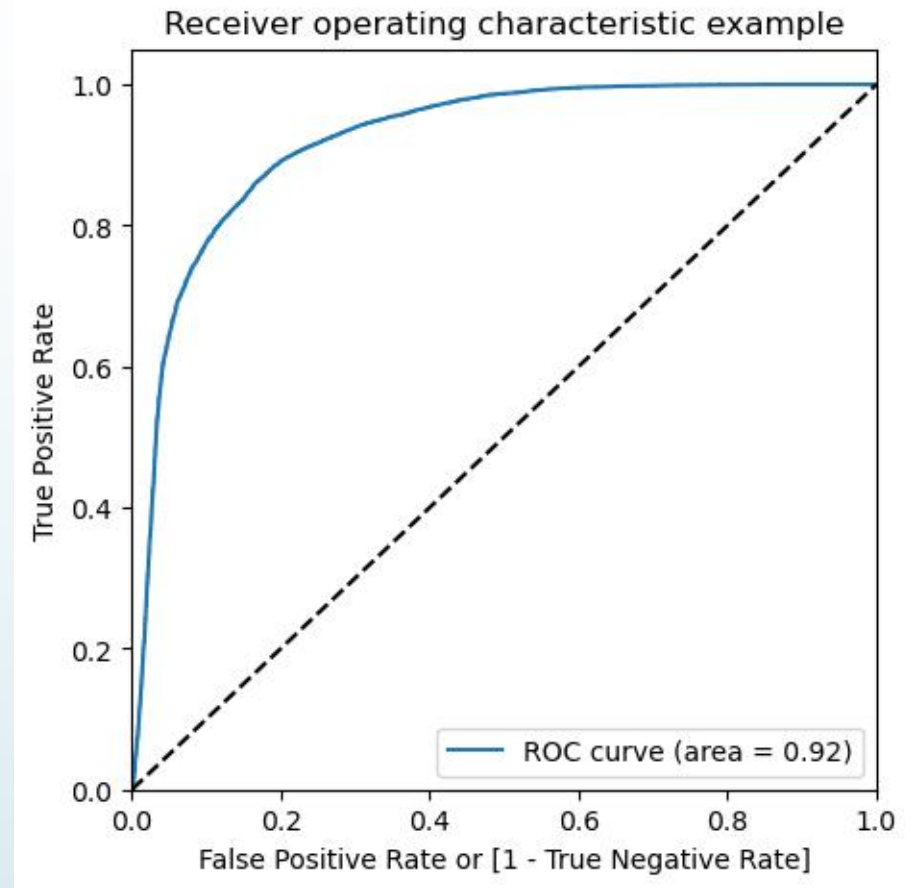
At point **0.6** where the three parameters cut each other, we can see that there is a balance between sensitivity and specificity with a good accuracy.

Here we are intended to achieve better sensitivity than accuracy and specificity. Though as per the above curve, we should take 0.6 as the optimum probability cutoff, we are taking **0.5** for achieving higher sensitivity, which is our main goal.



The 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.



Final conclusion with PCA

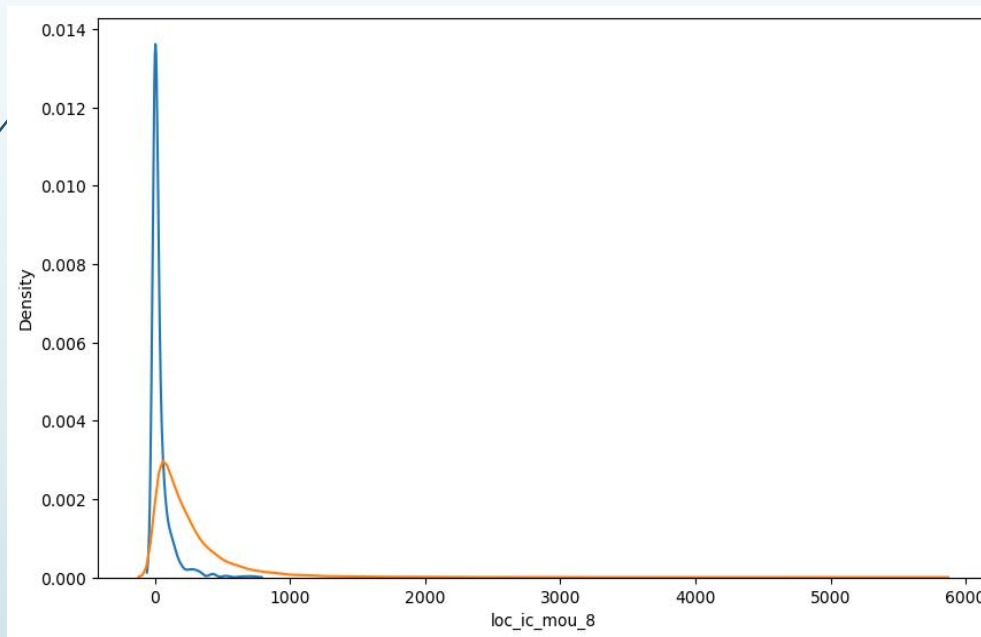
After trying several models we can see that for achieving the best sensitivity, which was our ultimate goal, the classic Logistic regression or the SVM models performs well. For both the models the sensitivity was approx. 81%. Also we have good accuracy of approx. 85%.

Final conclusion with no PCA

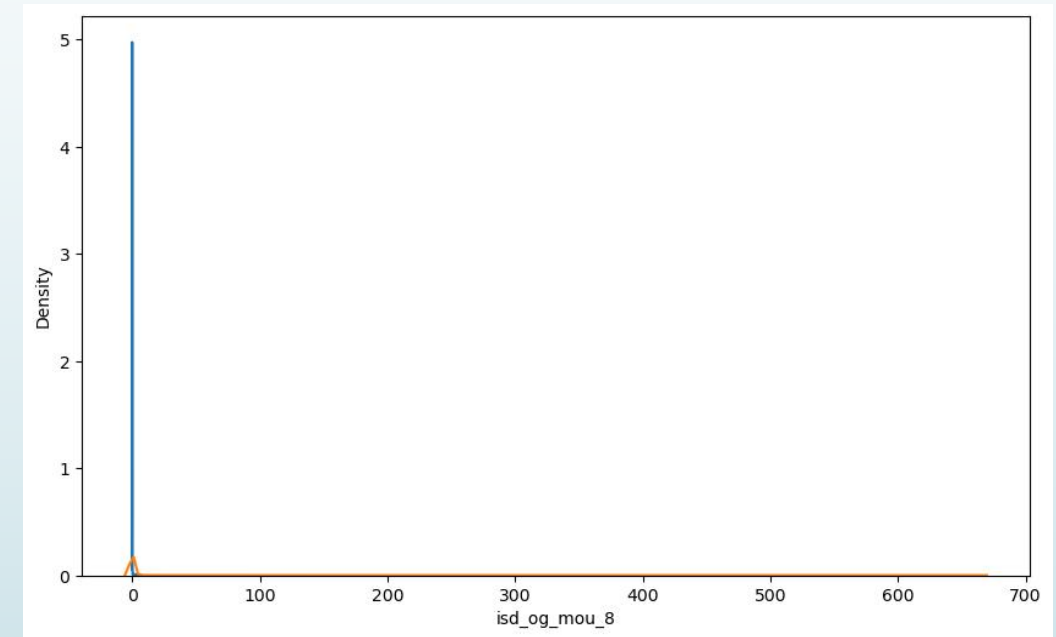
We can see that the logistic model with no PCA has good sensitivity and accuracy, which are comparable to the models with PCA. So, we can go for the more simplistic model such as logistic regression with PCA as it explains the important predictor variables as well as the significance of each variable.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

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 is mostly populated on the lower side than the non churn customers.

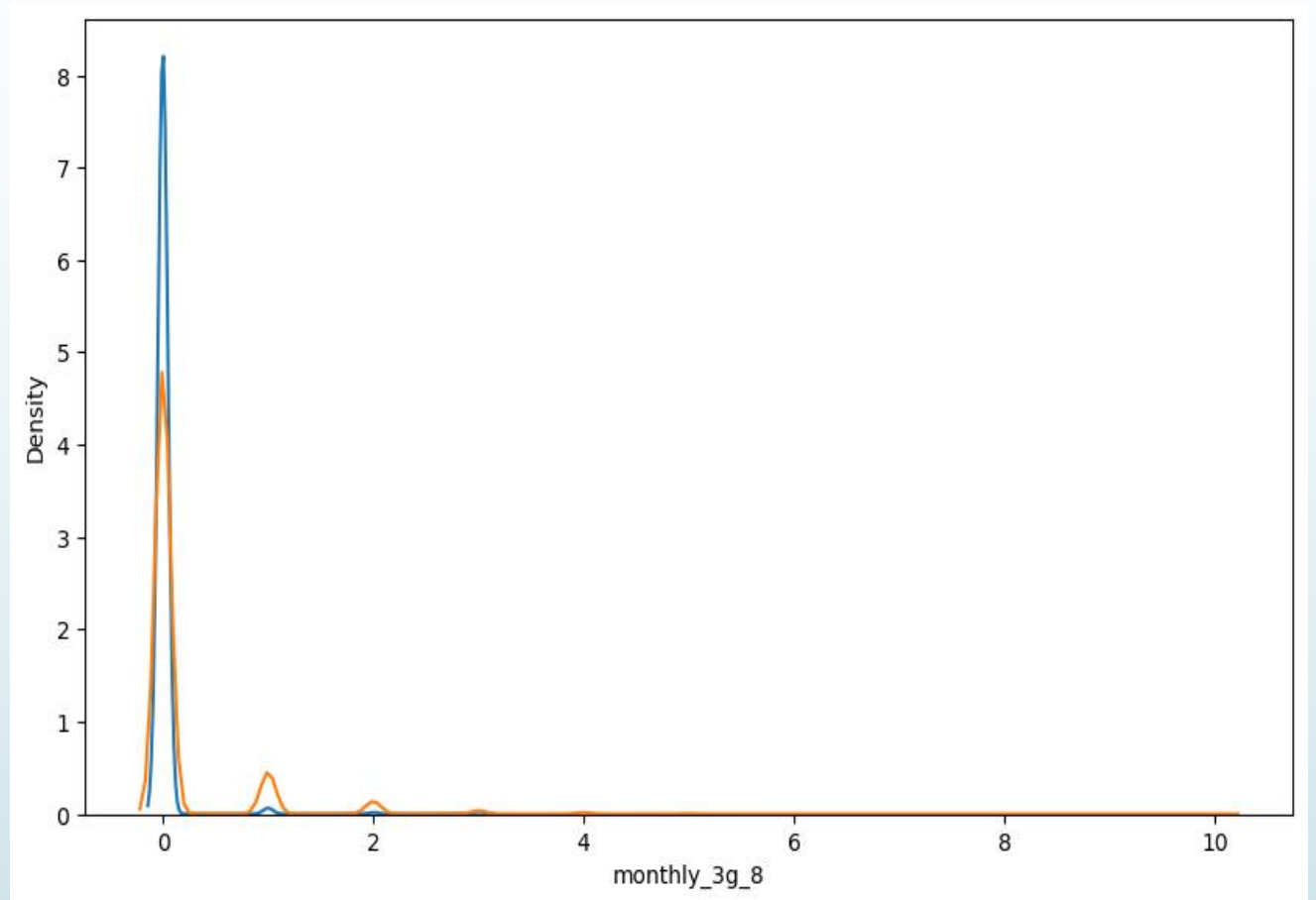


We can see that the ISD outgoing minutes of usage for the month of August for churn customers is dense approximately to zero. On the other hand for the non churn customers it is little more than the churn customers.



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 are spread across various numbers.



Business Recommendations

Top predictors :

- Besides are few top variables selected in the logistic regression model.
- 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) is lesser in the month of August than any other month, then there is a higher chance that the customer is likely to churn.

Variables	Coefficients
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

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, whose 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.

Conclusion

We ran the following models and have summarized our findings in the table below -

- Principle component Analysis and Regression
 - Logistic Regression with RFE and VIF
 - Decision Tree
 - Random Forest
- We see that almost on all models the values are coming very similar to each other and more often than not there is a trade off between sensitivity & specificity .

Since both the metrics are important, we feel that going ahead with Random forests.

We notice that the following 5 factors affect the churn rate considerably -

- 1) Total Incoming Minutes of usage in the August
- 2) Total Incoming Minutes of usage in the July
- 3) 2G data pack
- 4) Roaming
- 5) Sachet 2G

Also these metrics are inversely proportion to churn which means that we need to come up with campaigns that would keep people engaged either via calls (incoming) or on internet. One interesting thing to note here is that we see that a lot of people are hooked on 2G and hence are not churning. It presents us with a great opportunity that if shift these people from 2g to 3g then we have a greater chance of these people not churning. Hence discounts on 3G pack can be one of the popular marketing campaigns.