TELECOM CHURN CASE STUDY

IIIT-BANGALORE(B.A)

SUBMITTED BY: VYSHNAVI ACHAR H.V

VIMAL KANT

ABDUL WAHIDH MC

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

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

Definitions of churn

There are various ways to define churn, such as:

• **Revenue-based churn:** Customers who have not utilized any revenue-generating facilities such as mobile internet, outgoing calls, SMS etc. over a given period of time. One could also use aggregate metrics such as ‘customers who have generated less than INR 4 per month in total/average/median revenue. The main shortcoming of this definition is that there are customers who only receive calls/SMSes from their wageearning counterparts, i.e. they don’t generate revenue but use the services. For example, many users in rural areas only receive calls from their wage-earning siblings in urban areas.

• **Usage-based churn:** Customers who have not done any usage, either incoming or outgoing - in terms of calls, internet etc. over a period of time. A potential shortcoming of this definition is that when the customer has stopped using the services for a while, it may be too late to take any corrective actions to retain them. For e.g., if you define churn based on a ‘two-months zero usage’ period, predicting churn could be useless since by that time the customer would have already switched to another operator.

- **High-value churn:** 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.In this project, you will define high-value customers based on a certain metric (mentioned later below) and predict churn only on high-value customers

STEPS INVOLVED DURING TELECOM CHURN CASE ANALYSIS

1. Reading, understanding and visualizing the data
2. Preparing the data for modeling
3. Building the model
4. Evaluate the model

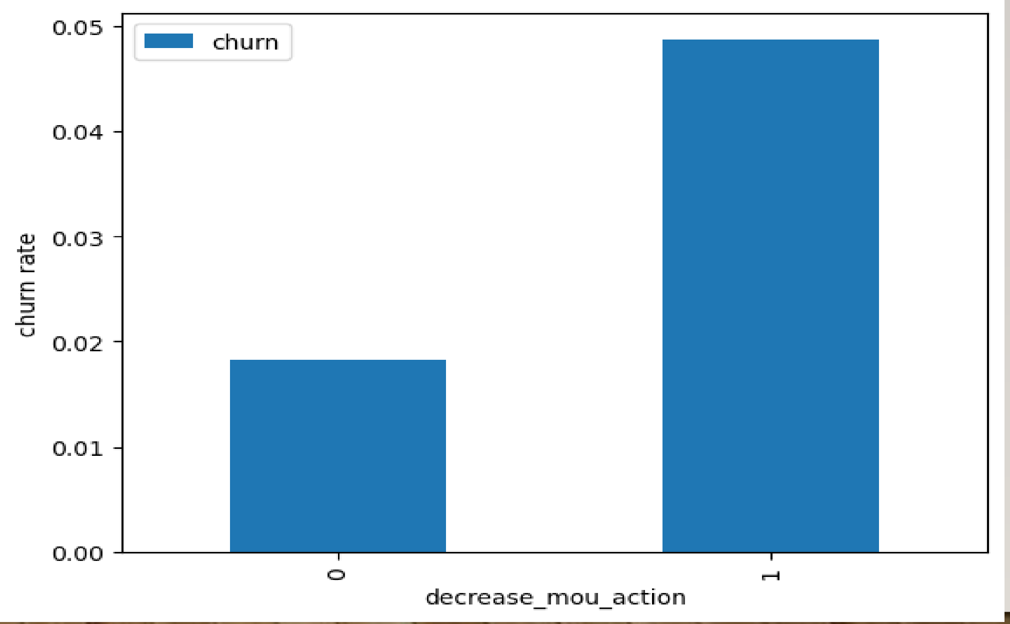
1. Reading, understanding and visualizing the data

- This file contains 99999 rows and 226 columns.
- After deleting all missing values rows and unnecessary columns, there are 27991 rows and 178 columns left. We have lost almost 7% records. But we have enough number of records to do our analysis.
- 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.
- After deleting all the attributes corresponding to the churn phase, It came out 3.39 %of churn rate are there

Exploratory Data Analysis

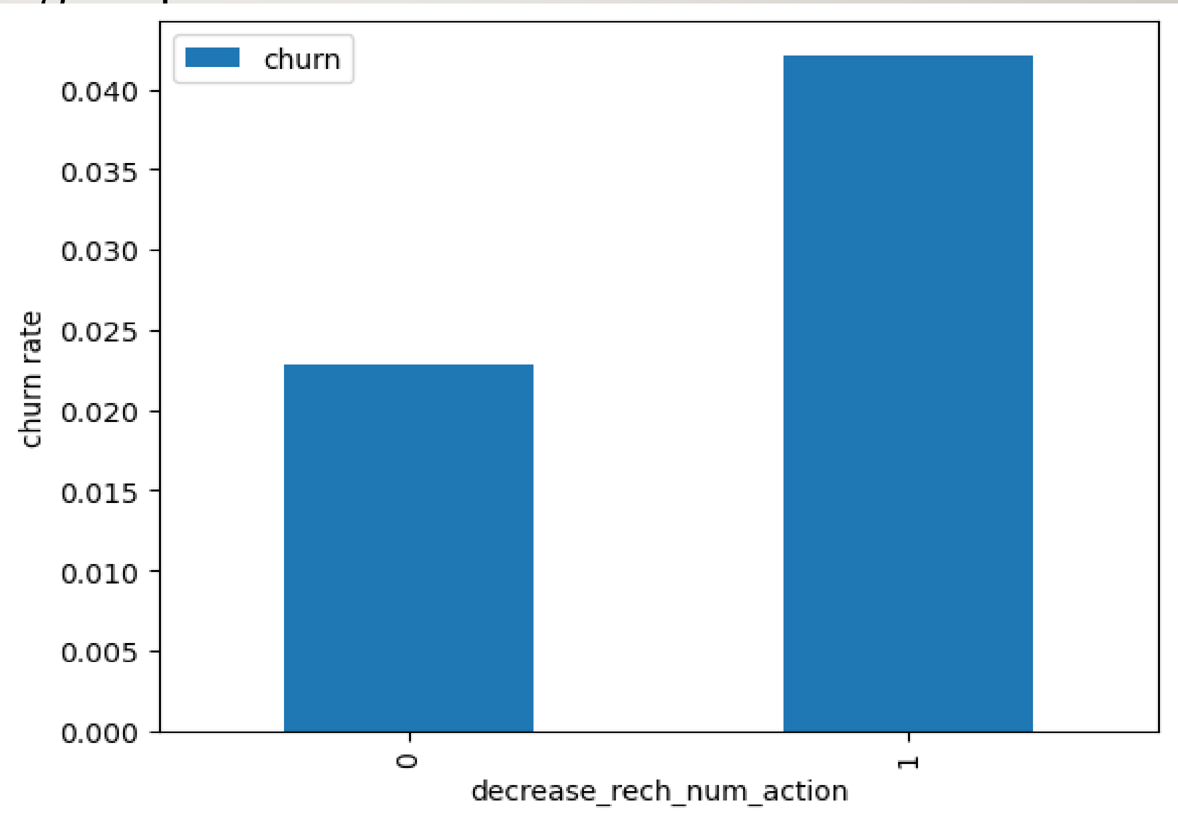
❖ Univariate analysis

- 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.



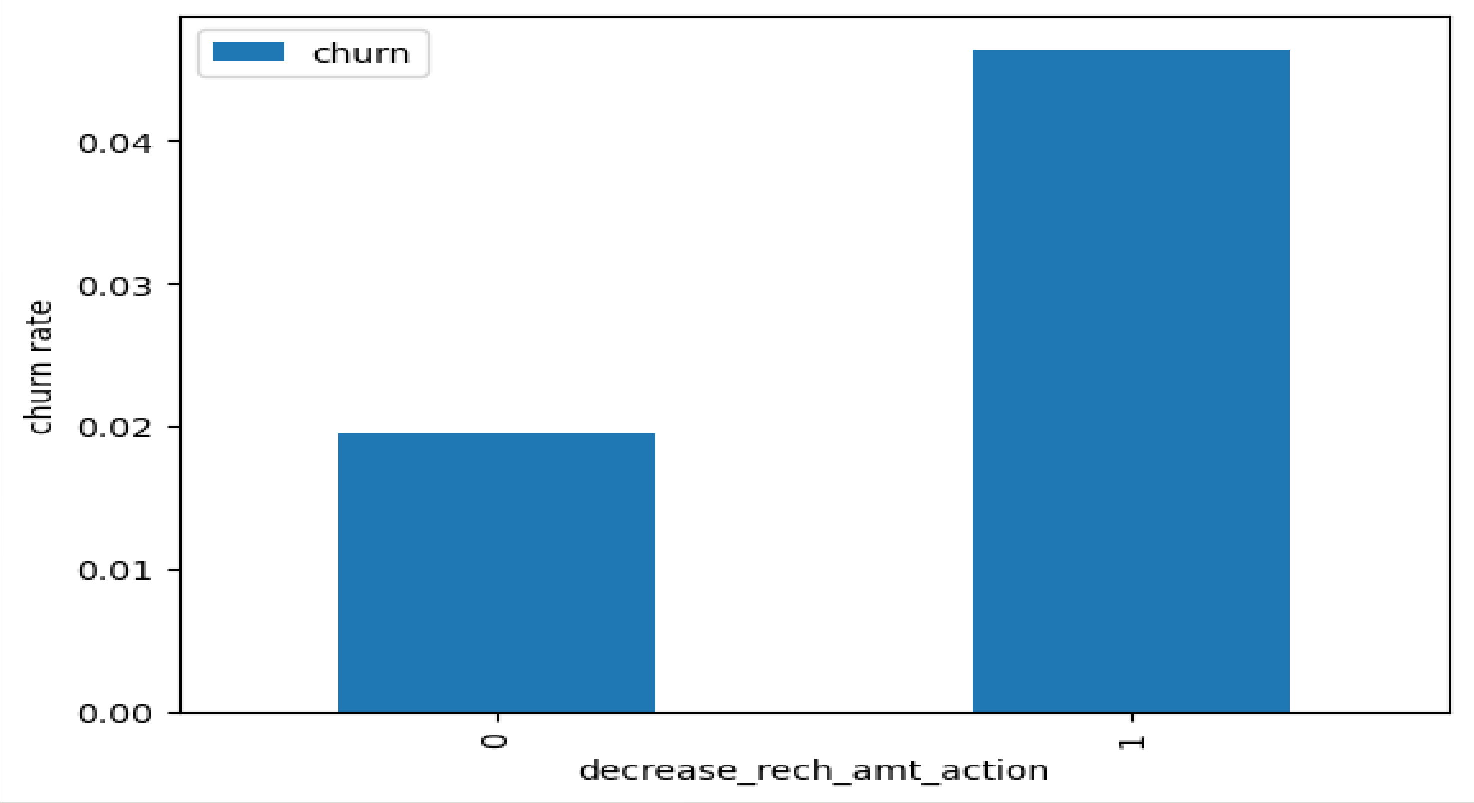
- **Churn rate on the basis whether the customer decreased her/his number of recharge in action month.**

- 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



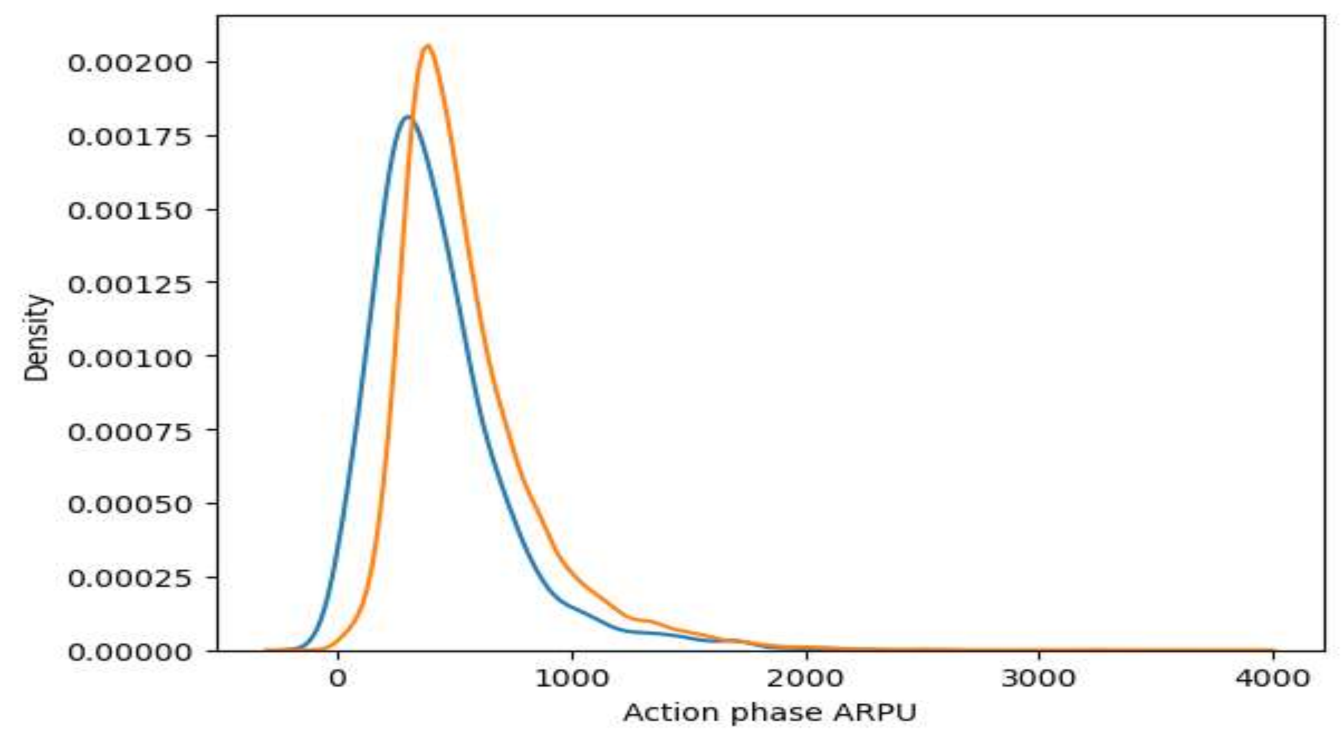
- **Churn rate on the basis whether the customer decreased her/his volume based cost in action month.**

- 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



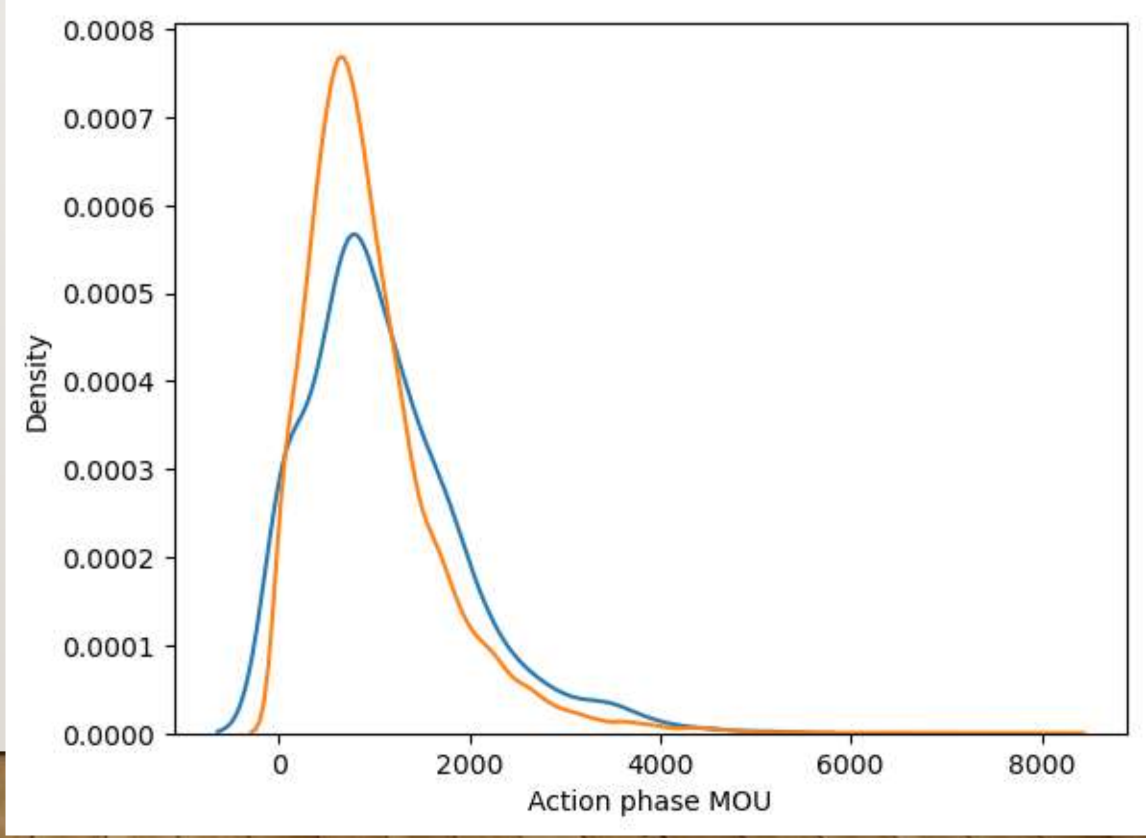
Analysis of the average revenue per customer (churn and not churn) in the action phase.

- 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



Analysis of the minutes of usage MOU (churn and not churn) in the action phase.

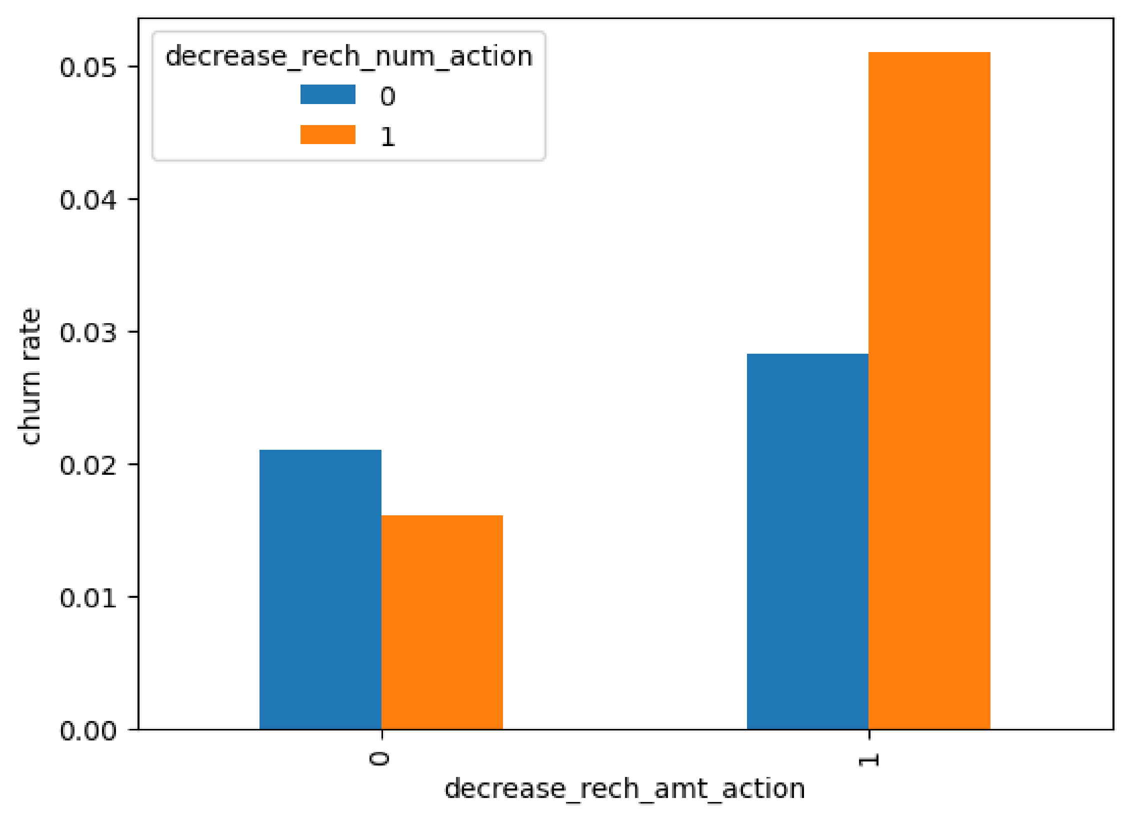
- Minutes of usage(MOU) of the churn customers is mostly populated on the 0 to 2500 range. Higher the MOU, lesser the churn probability



Bivariate analysis

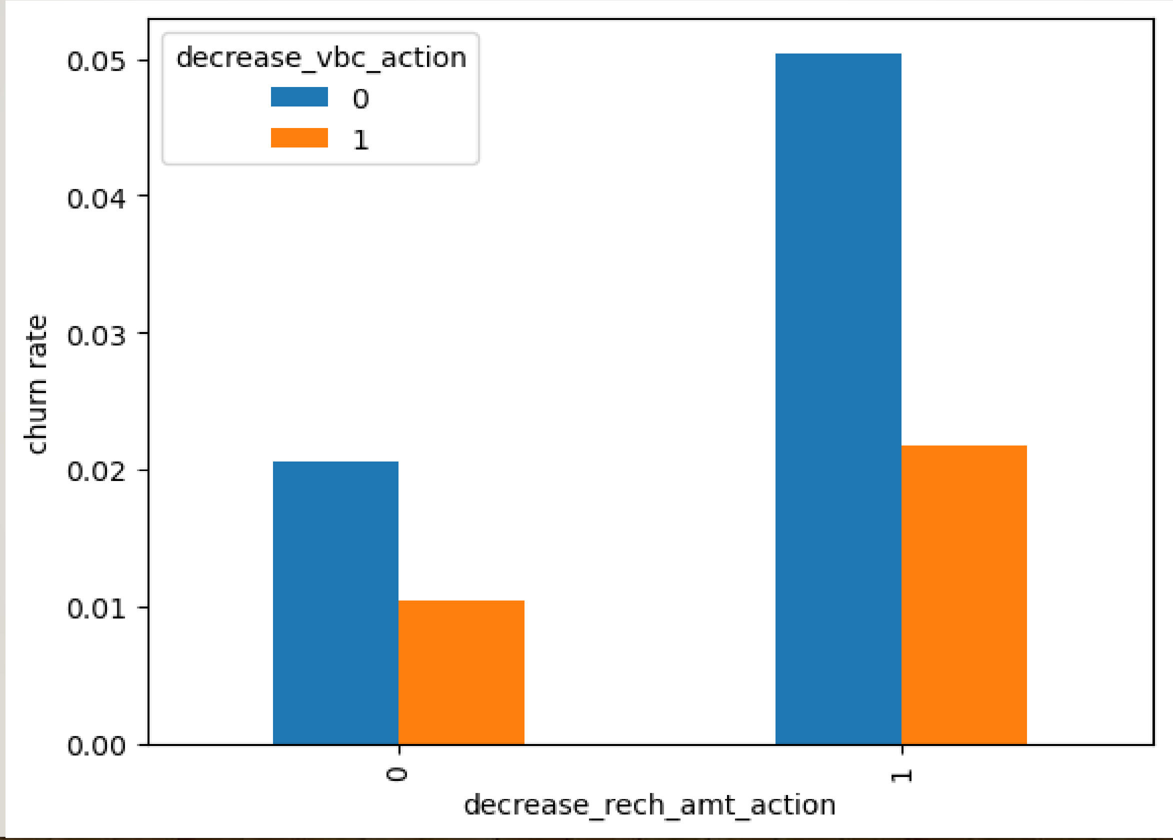
Analysis of churn rate by the decreasing recharge amount and number of recharge in the action phase.

- 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



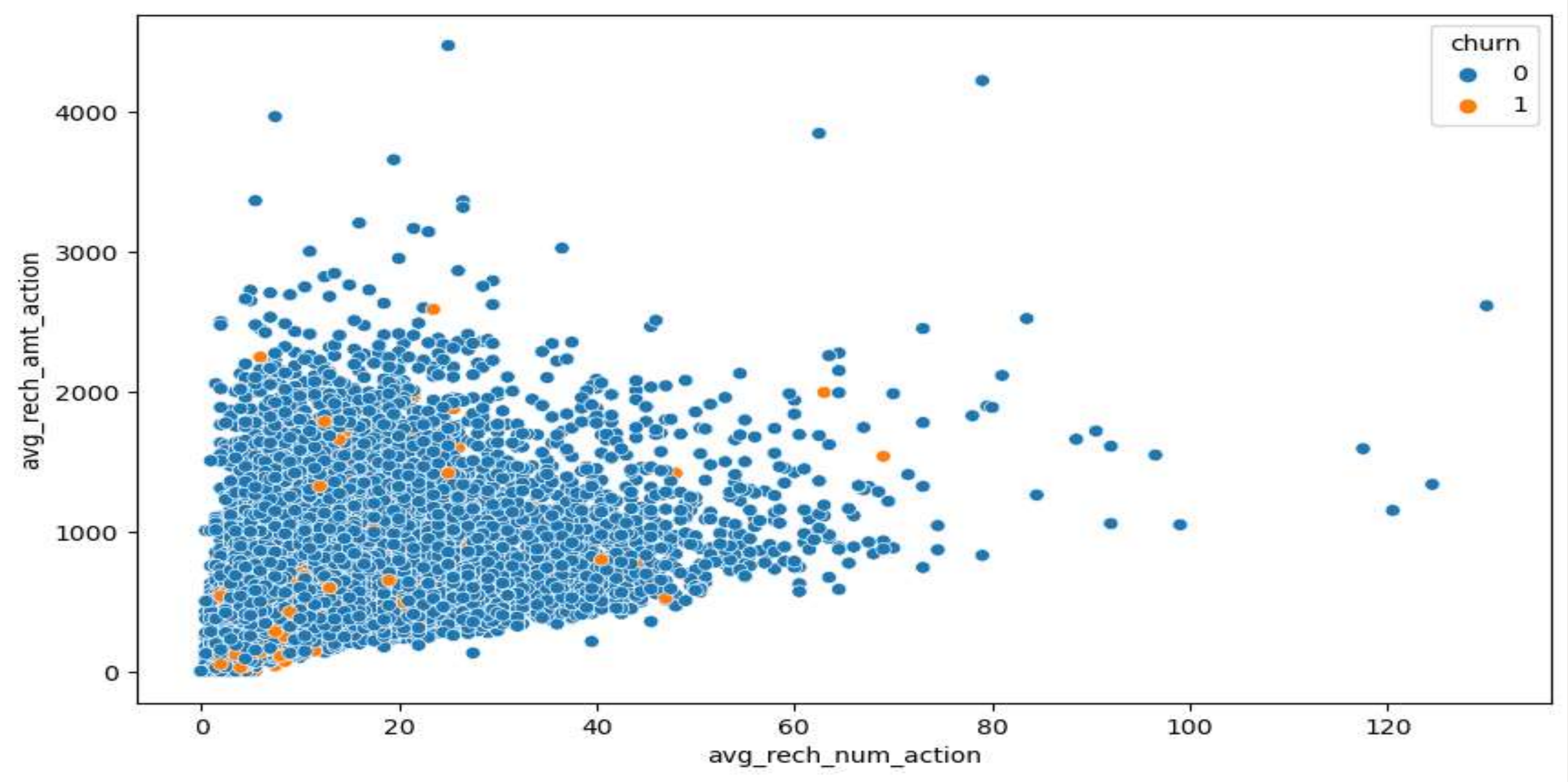
Analysis of churn rate by the decreasing recharge amount and volume based cost in the action phase.

- 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



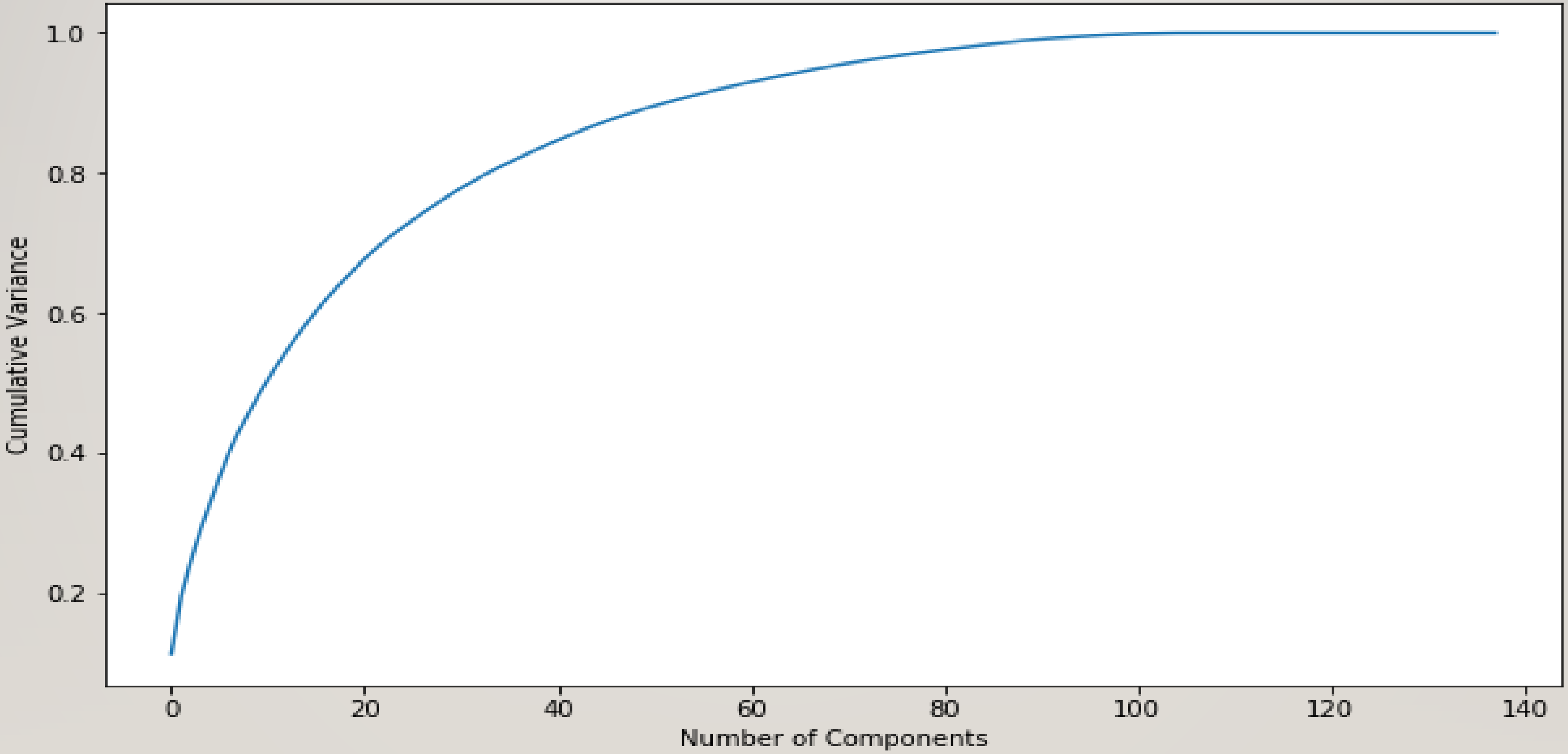
Analysis of recharge amount and number of recharge in action month.

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



Model with PCA

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



Logistic regression with PCA

- **Model summary**

- **Train set**

  Accuracy = 0.86

  Sensitivity = 0.89

  Specificity = 0.83

- **Test set**

  Accuracy = 0.83

  Sensitivity = 0.81

  Specificity = 0.83

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

Support Vector Machine(SVM) with PCA

❖ Model summary

- **Train set**

  Accuracy = 0.89
  Sensitivity = 0.92
  Specificity = 0.85

- **Test set**

  Accuracy = 0.85
  Sensitivity = 0.81
  Specificity = 0.85

Decision tree with PCA

- **Model summary**

- **Train set**

  Accuracy = 0.90
  Sensitivity = 0.91
  Specificity = 0.88

- **Test set**

  Accuracy = 0.86
  Sensitivity = 0.70
  Specificity = 0.87

We can see from the model performance that the Sensitivity has been decreased while evaluating the model on the test set. However, the accuracy and specificity is quite good in the test set

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 preforms well. For both the models the sensitivity was approximate 81%. Also we have good accuracy of approximate 85%

Model With No PCA

Random forest with No PCA.

❖ **Model summary**

- **Train set**

	precision	recall	f1-score	support
0	0.99	0.97	0.98	5348
1	0.42	0.60	0.49	193
accuracy			0.96	5541
Macro avg	0.70	0.79	0.73	5541
Weighted avg	0.97	0.96	0.96	5541

- **Test set**

  Accuracy = 0.95
  Sensitivity = 0.60
  Specificity = 0.96

The model is able to predict the data with almost 95% accuracy on test data which is a good score. The model is also not underfitting

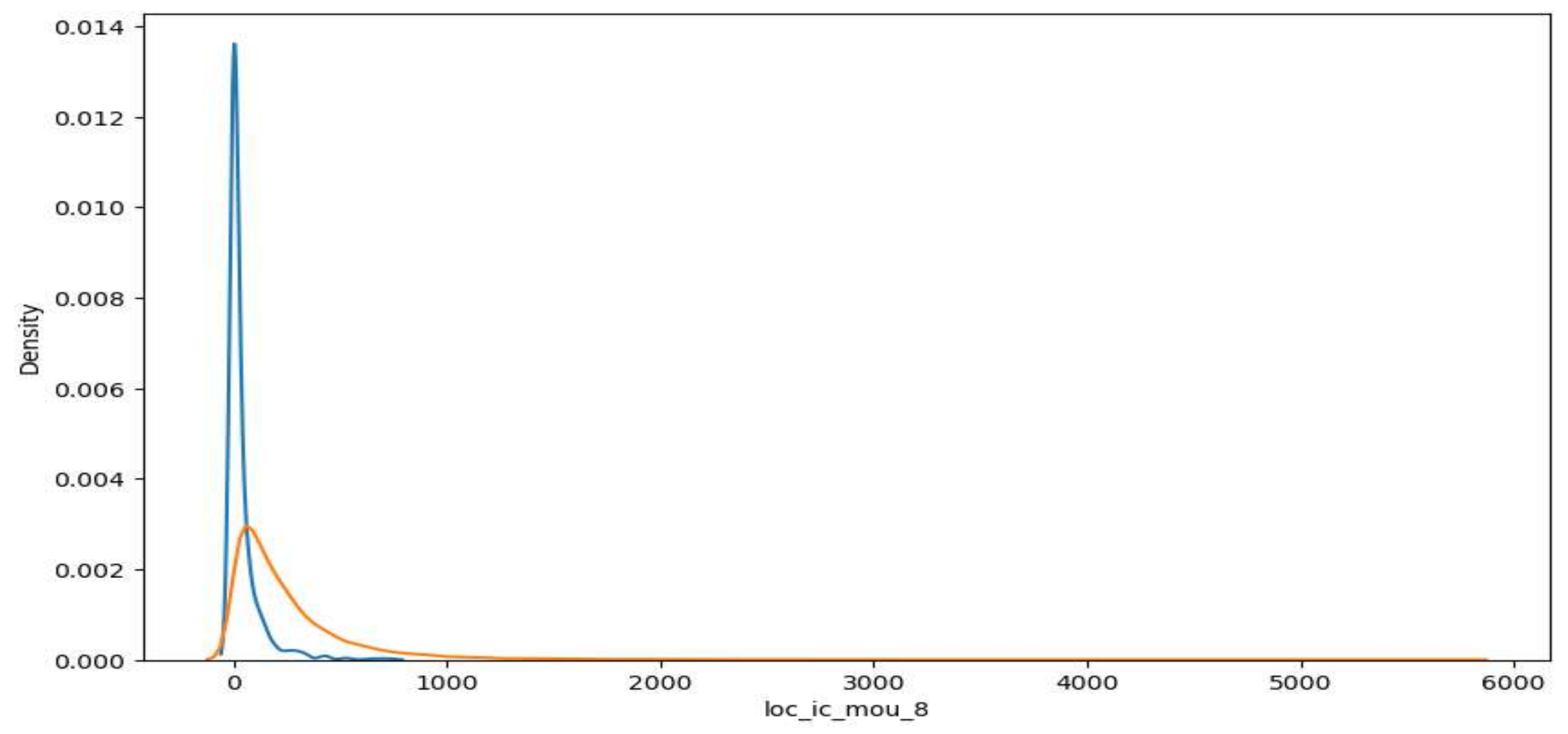
Final conclusion with No PCA

- We can see that the random forest model with no PCA has good Specificity 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

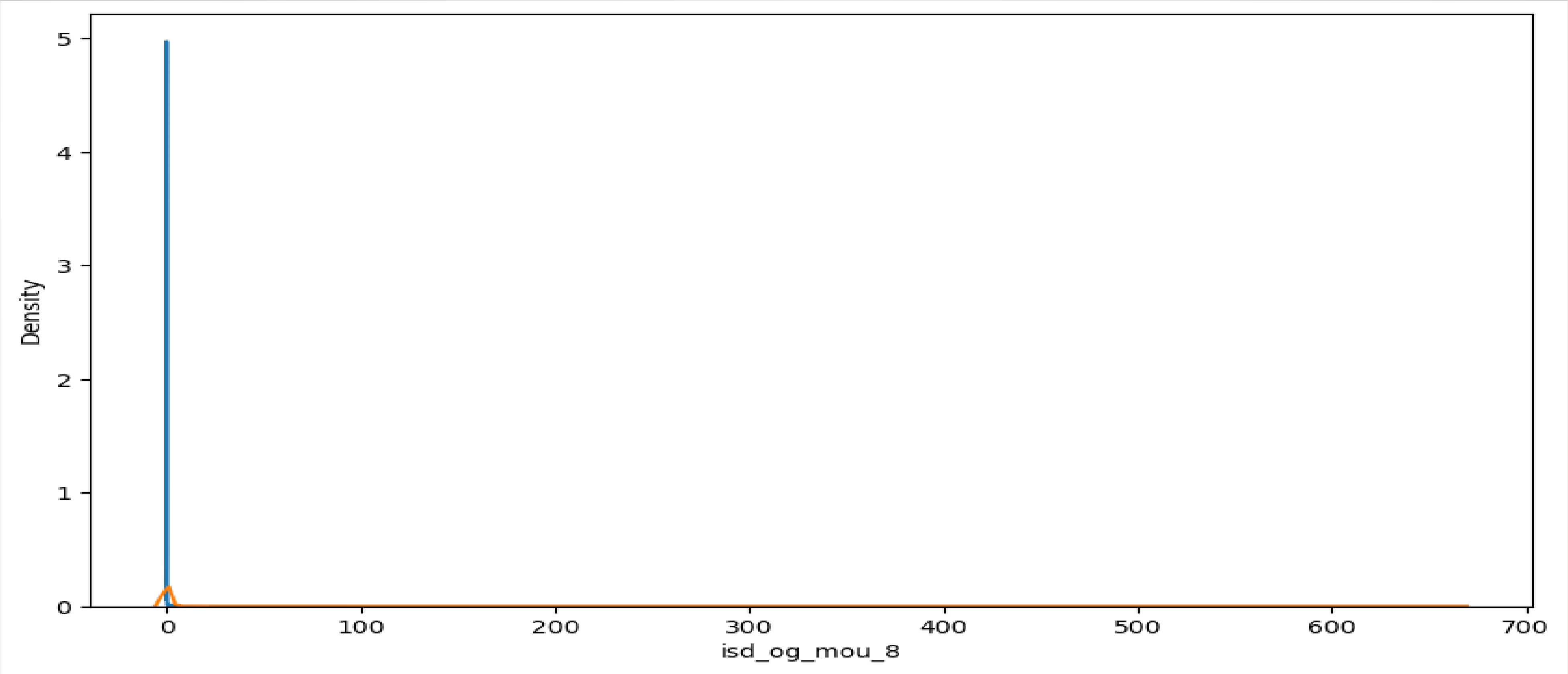
Plots of important predictors for churn and non churn customers

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



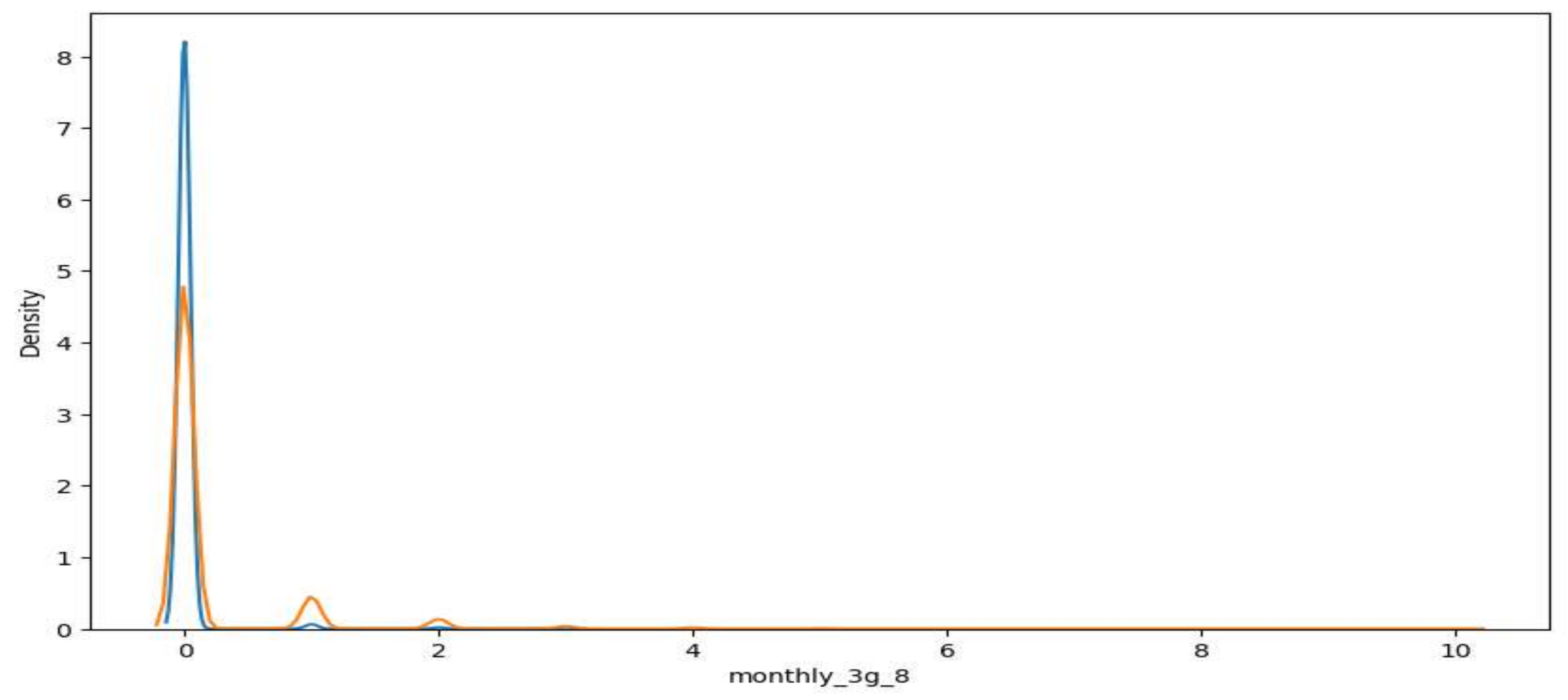
Plotting isd_og_mou_8 predictor for churn and not churn customers

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



Plotting monthly_3g_8 predictor for churn and not 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 variables, which have higher coefficients, churn distribution.



The summary of the models built using permutations of the above learning algorithms, dataset type and Tuning types is displayed below. It show the Confusion Matrix output of each Model and the Sensitivity score

Models	**Confusion matrix**		**Sensitivity**
Logistic Regression- With PCA-With Parameter Tuning	4452	896	0.8134715025906736
	36	157	
Random forest Without PCA	5185	163	0.6010362694300518
	77	116	
Support Vector Machine(SVM) with PCA-PCA-WITH PARAMETER TUNING	4557	791	0.8134715025906736
	36	157	
Decision tree with PCA-With Parameter Tuning	4632	716	0.6994818652849741
	58	135	

Business 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 That means or the customers, whose roaming outgoing minutes of usage is increasing are more likely to churn

Summary - Results, Findings and Inferences

Following were the learning algorithms used

- ❖ Logistic Regression
- ❖ Decision Tree
- ❖ Support Vector Machine(SVM)
- ❖ Random Forest

Following were the type of data set combinations used

- ❖ With Dimensionalty Reduction (i.e. PCA)
- ❖ Without Dimensionalty Reduction (i.e. without PCA)

Also, Following were the Tuning made

- ❖ With Hyperparameter Tuning
- ❖ Without Hyperparameter Tuning

- It could be easily inferred that the best model derived was using Logistic Regression on dataset reduced dimensionally using PCA
- The key parameter to measure here was Sensitivity as we had to measure the customers who churned correctly not predicting those who haven't churned
- The best model derived had Sensitivity ~0.81 and Accuracy ~0.83 which was reasonable. On further trying to improve the Sensitivity the Accuracy of the model was taking a hit. Therefore we reached an optimal level
- A separate model was built using Logistic Regression to identify the predictor variables. Based on that following key variables were identified:
    1) Total outgoing calls in minutes of usages of voice calls for the 7th Month.
    2) Incoming Local voice calls - within same telecom circle between Operator T to other operator mobile in terms of Minutes of usage for 8th Month.
    3) Average mobile internet (3g) usage by volume (in MB) for 8th Month.
    4) Volume based cost - when no specific scheme is not purchased and paid as per usage; the volume based cost for 8th Month.
- On further analyzing each of these 4 variables across the 3 months i.e. month 6,7,8 both the Good Phase months and the Action Phase month the following could easily be derived:
    1) Total outgoing calls in minutes of usages of voice calls keeps on decreasing for churn customers, whereas for non-churn customers it broadly remains the same.

2) The Local incoming calls in minutes of usages of voice calls from other operators' mobile networks keeps on decreasing for churn customers, whereas for non-churn customers it broadly remains the same.

3) The Average mobile internet (3g) usage by volume (in MB) keeps on decreasing for churn customers, whereas for non-churn customers it broadly remains the same.

4)The volume based cost keeps on decreasing for churn customers, whereas for non-churn customers it broadly remains the same.

❖ As on all the parameters the trends are going down i.e. the total outgoing voice Valls, local incoming from mobile, average mobile internet usage by volume it appears that the customers are largely not using the SIM card or may have placed it in the secondary slot of a dual SIM. Therefore, this could only mean that the customers are not happy with the service provided by the telecom company or a competitor is offering better services / offers. However, given the study decline noticed, it appears that the customers are not happy with the current services provided by the telecom company. Therefore, the company needs to

1) Offer some incentives as highlighted in point #1 above to keep the customers engaged and avoid them churning out

2) Improve its Network Capabilities to avoid call drops and data connectivity issues

3) Improve its Call Center Operations to ensure that the customer reported complaints are resolved as soon as possible

Recommended Strategies to manage Customer Churn

Based on the key predictor variables identified above following are some of the strategies that could be adopted by the telecom company to reduce the churn

- ❖ If there is a drop in the total outgoing calls in the "Good Phase" i.e. between month 6 and 7 then provide some incentives to the customer in the "Action Phase" i.e. month 8 which could like some of the following
    1) Offering free outgoing calls say for a month. This will ensure that for the same balance the customer will stick for at least a month now.
    2) Offer healthy Cashback amount on recharge for the 8th and 9th month together.
    3) Free data pack on recharge for month 8 and 9 together.

- ❖ As we are dealing with High Value Customers here, most of them should be able to pay for two months recharge together in advance in exchange of some good offers. Also, since acquiring a customer costs 5-10 times more than retaining an one, these offers shouldn't harm the company's financials much.
- ❖ Lowering of Outgoing Calls could also suggest Network Congestion / Frequent call drops in a paricular area etc. i.e. some back end network related issues. The Telecom Company should analyze if all the churn customers belong to a particular micro-region of a City and if yes, then they will have to look at ways to lower the network issues. If it means like adding more towers etc. which could take more time, then some offers like mentioned above could help to engage the customer in the interim.

As on all the parameters the trends are going down i.e. the total outgoing voice calls, local incoming from mobile, average mobile internet usage by volume it appears that the customers are largely not using the SIM card or may have placed it in the secondary slot of a dual SIM. Therefore, this could only mean that the customers are not happy with the service provided by the telecom company or a competitor is offering better services / offers. However, given the study decline noticed, it appears that the customers are not happy with the current services provided by the telecom company. Therefore, the company needs to:

1) Offer some incentives as highlighted in point #1 above to keep the customers engaged and avoid them churning out.

2) Improve its Network Capabilities to avoid call drops and data connectivity issues.

3) Improve its Call Center Operations to ensure that the customer reported complaints are resolved as soon as possible