Telecom Churn Case Study

Submitted by: Ravindra Verma,
Ann, and Vishnu Mohandas Menon
DSC-43 (BA)
IIIT-Bangalore

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 wage-earning 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

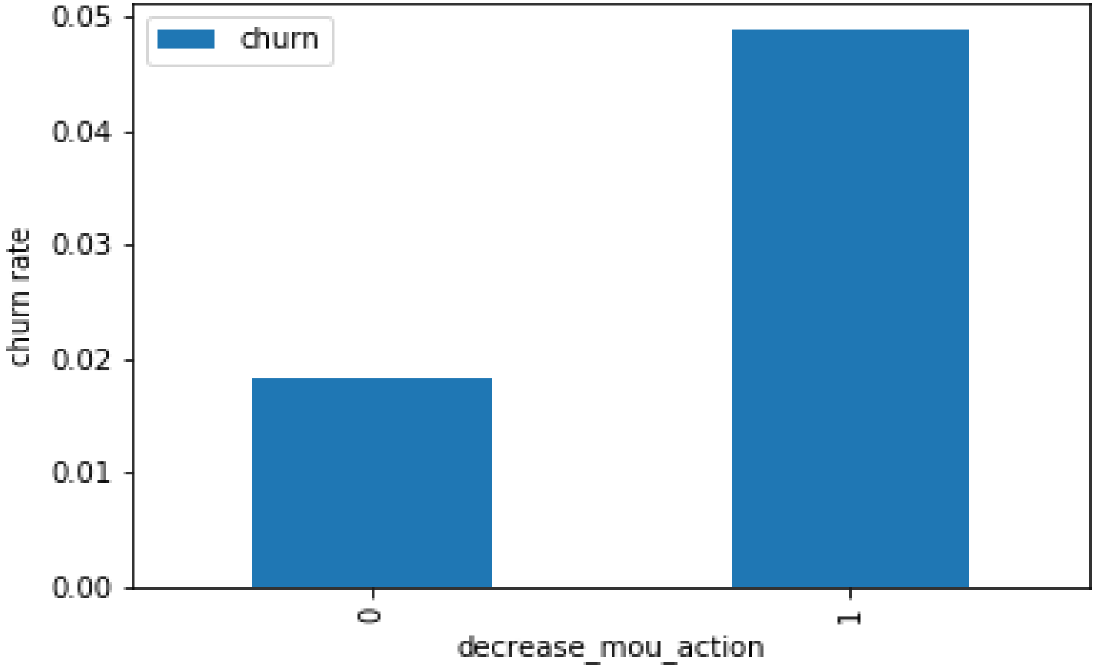
- 1. Reading, understanding and visualizing the data
- 2. Preparing the data for modeling
- 3. Building the model
- 4. Evaluate the model

1. Reading, understanding and visualizing the data

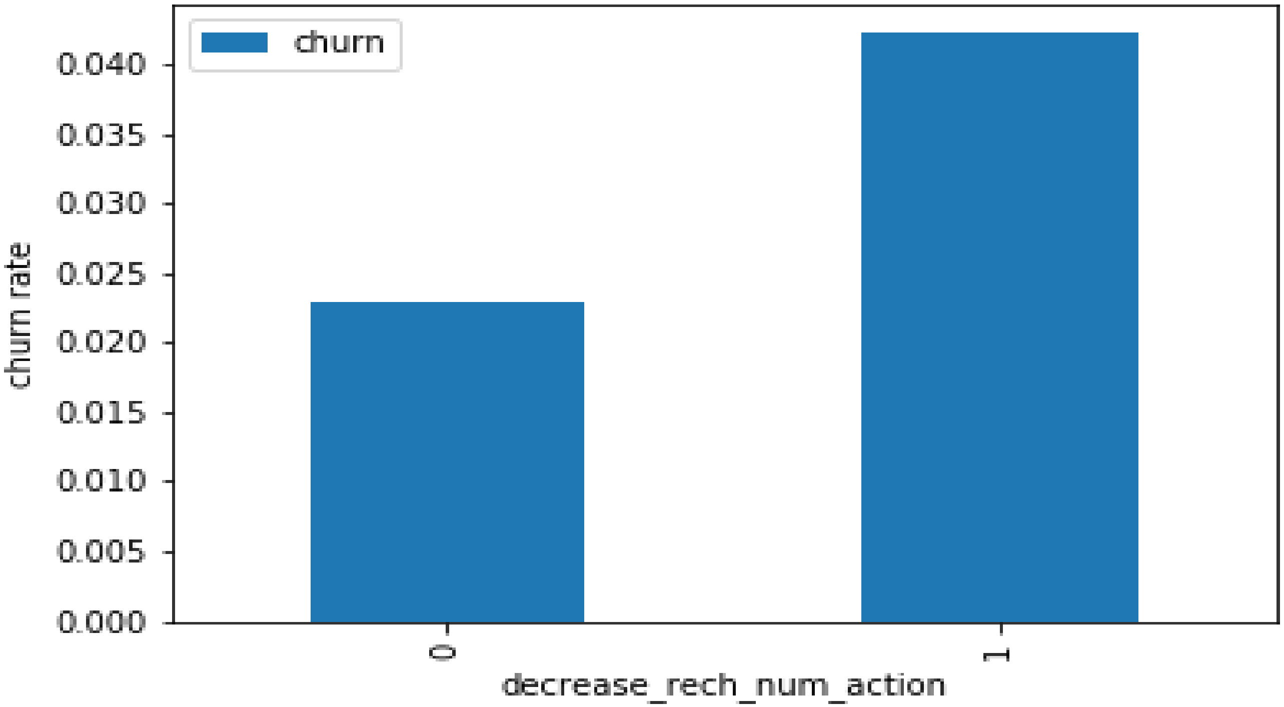
- This file contains 99999 rows and 226 columns.
- After deleting all missing values rows and unnecessary columns, there are 27991 rows and 178 columns left. We have lost almost 7% records. But we have enough number of records to do our analysis.
- Now tag the churned customers (churn=1, else 0) based 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

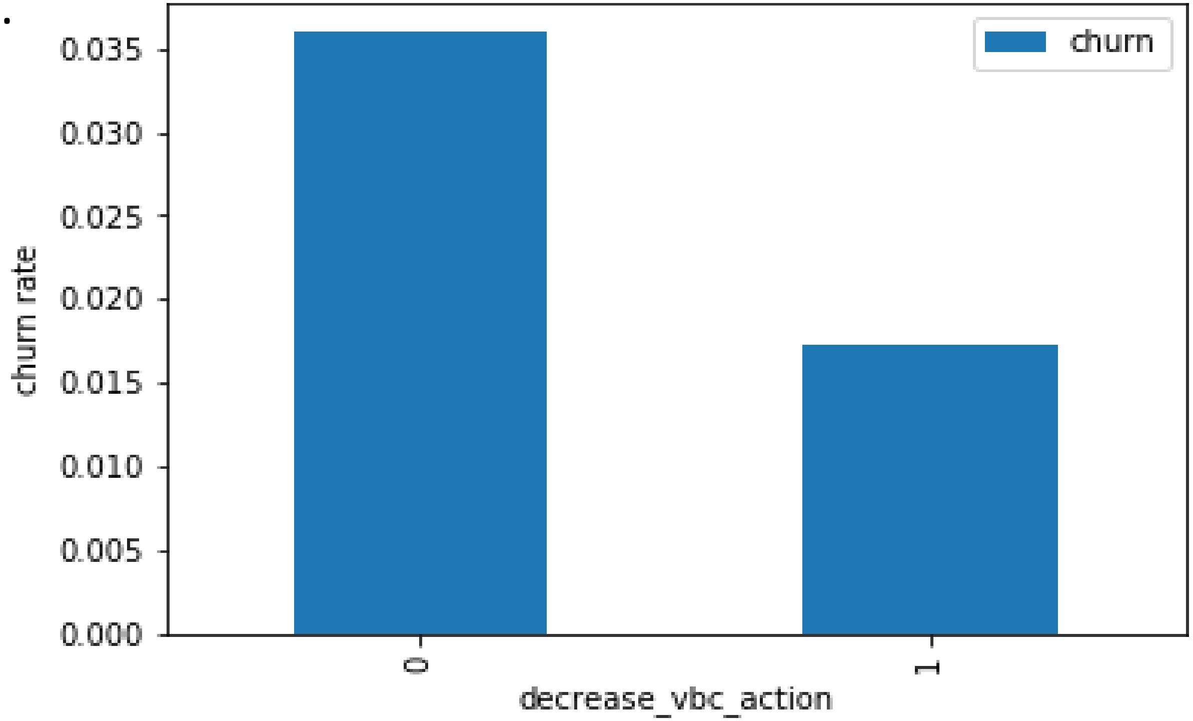
- **Univariate analysis**
- **Churn rate on the basis whether the customer decreased her/his MOU in action month.**
- We can see that the churn rate is more for the customers, whose minutes of usage(mou) decreased in the action phase than the good phase.



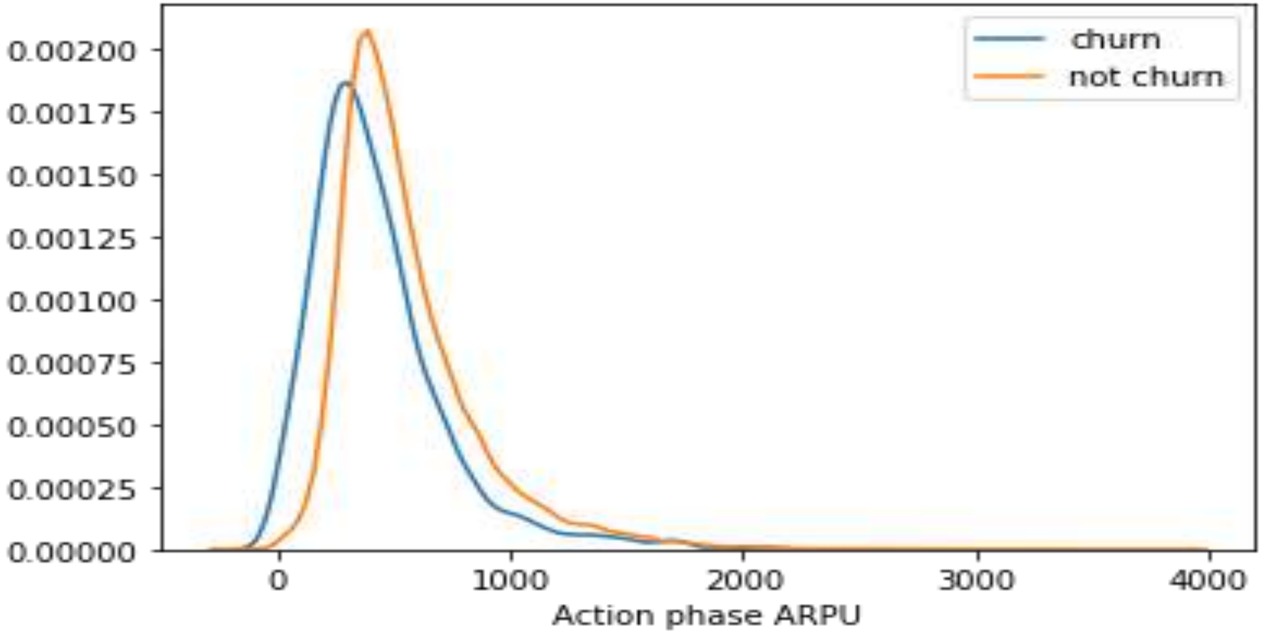
- **Churn rate on the basis whether the customer decreased her/his number of recharge in action month**.
- As expected, the churn rate is more for the customers, whose number of recharge in the action phase is lesser than the number in good phase.



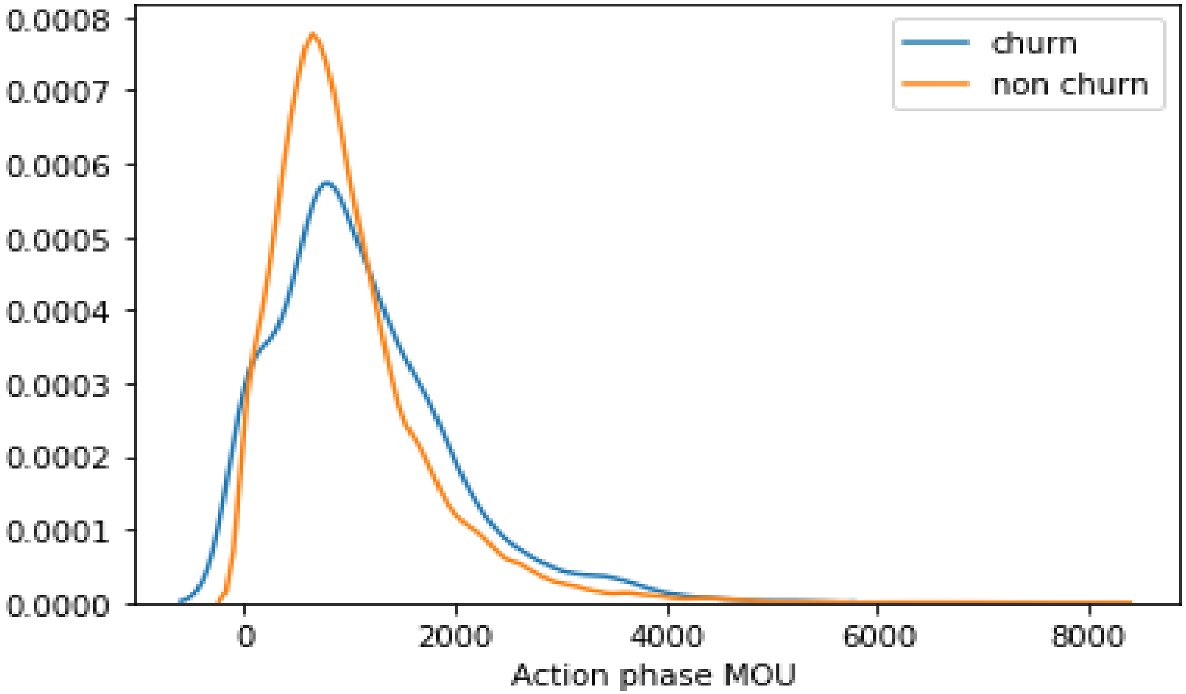
- **Churn rate on the basis whether the customer decreased her/his volume based cost in action month.**
- Here we see the expected result. The churn rate is more for the customers, whose volume based cost in action month is increased. That means the customers do not do the monthly recharge more when 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.

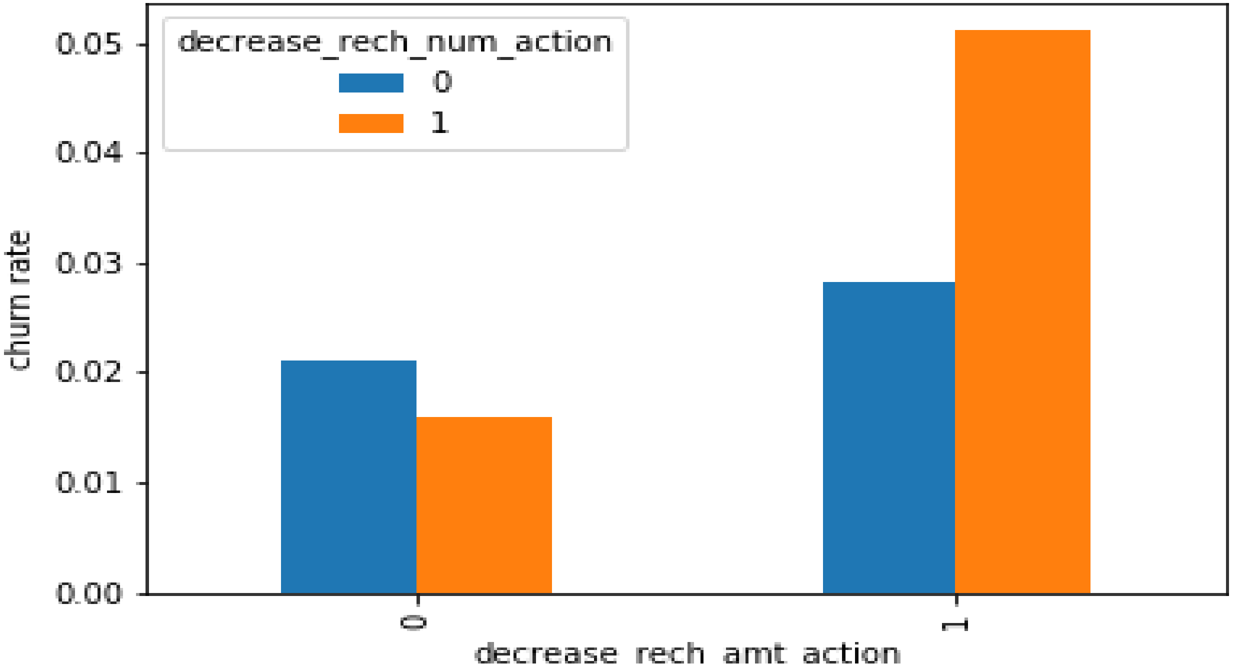


- ***Analysis of the minutes of usage MOU (churn and not churn) in the action phase.***
- Minutes of usage(MOU) of the churn customers is mostly populated on the 0 to 2500 range. Higher the MOU, lesser the churn probability.

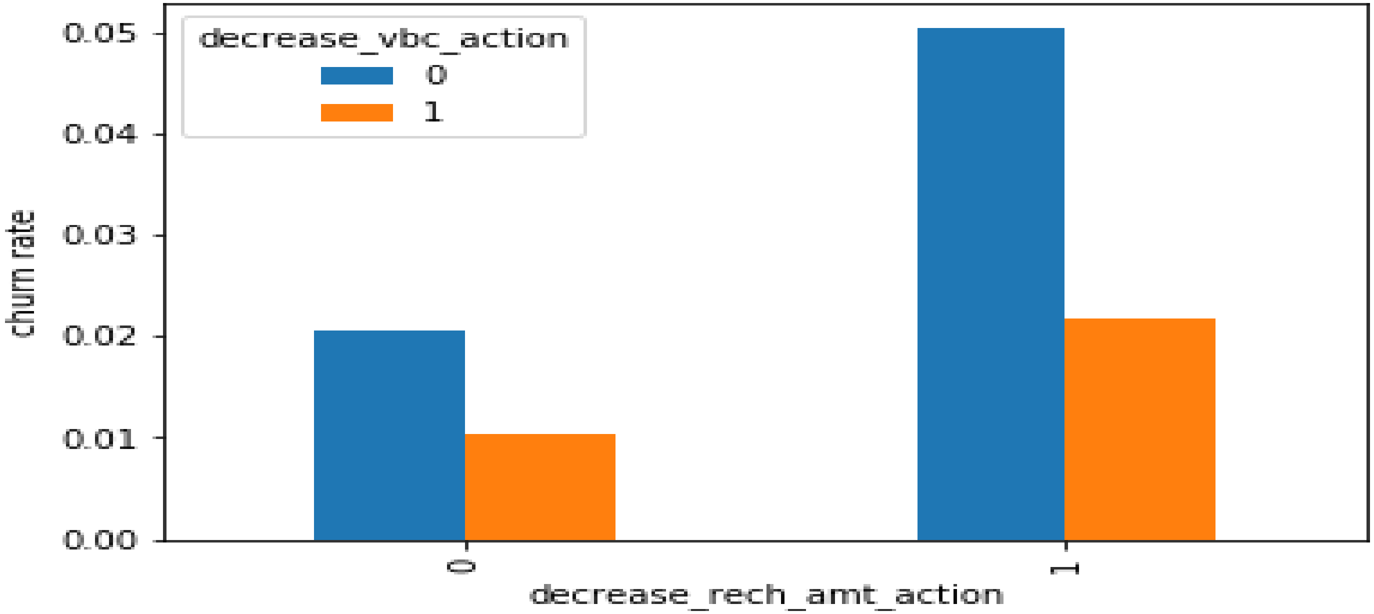


Bivariate analysis

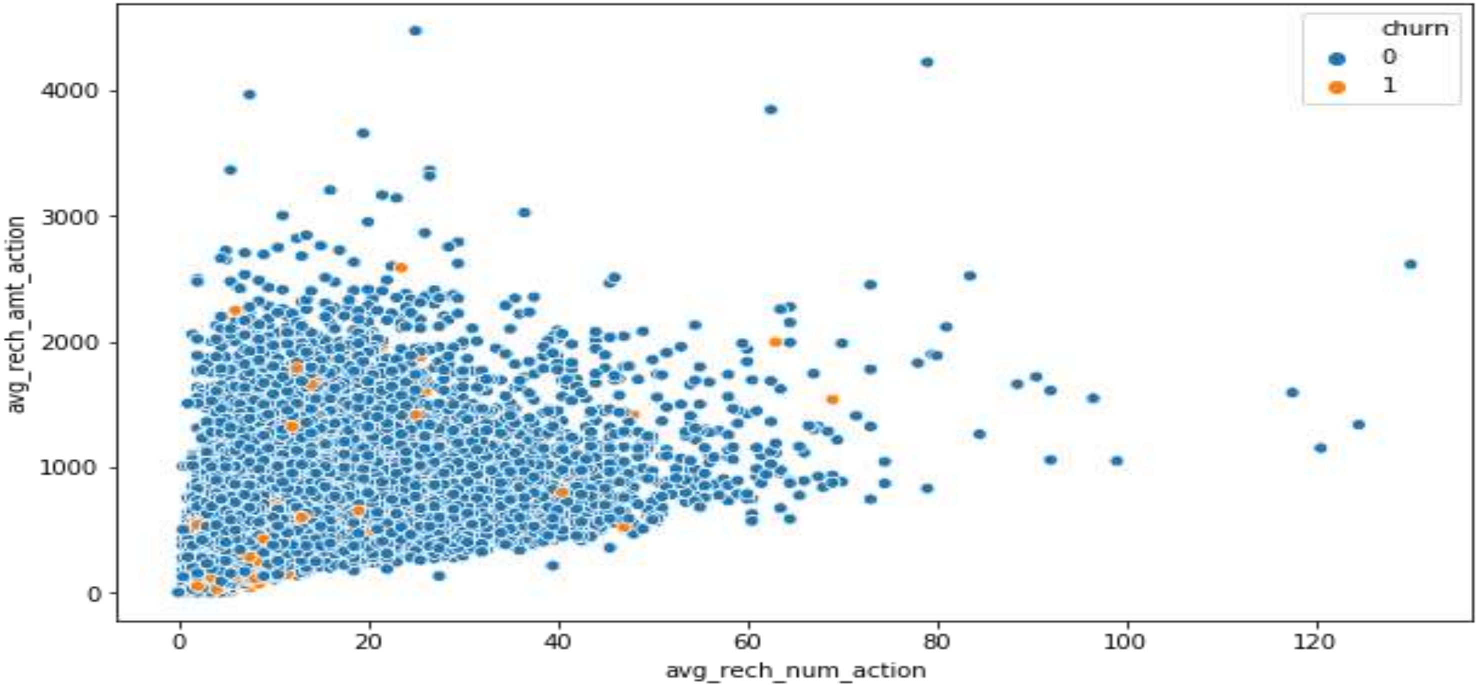
- ***Analysis of churn rate by the decreasing recharge amount and number of recharge in the action phase.***
- We can see from the above plot, that the churn rate is more for the customers, whose recharge amount as well as number of recharge have decreased in the action phase than the good phase.



- ***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.

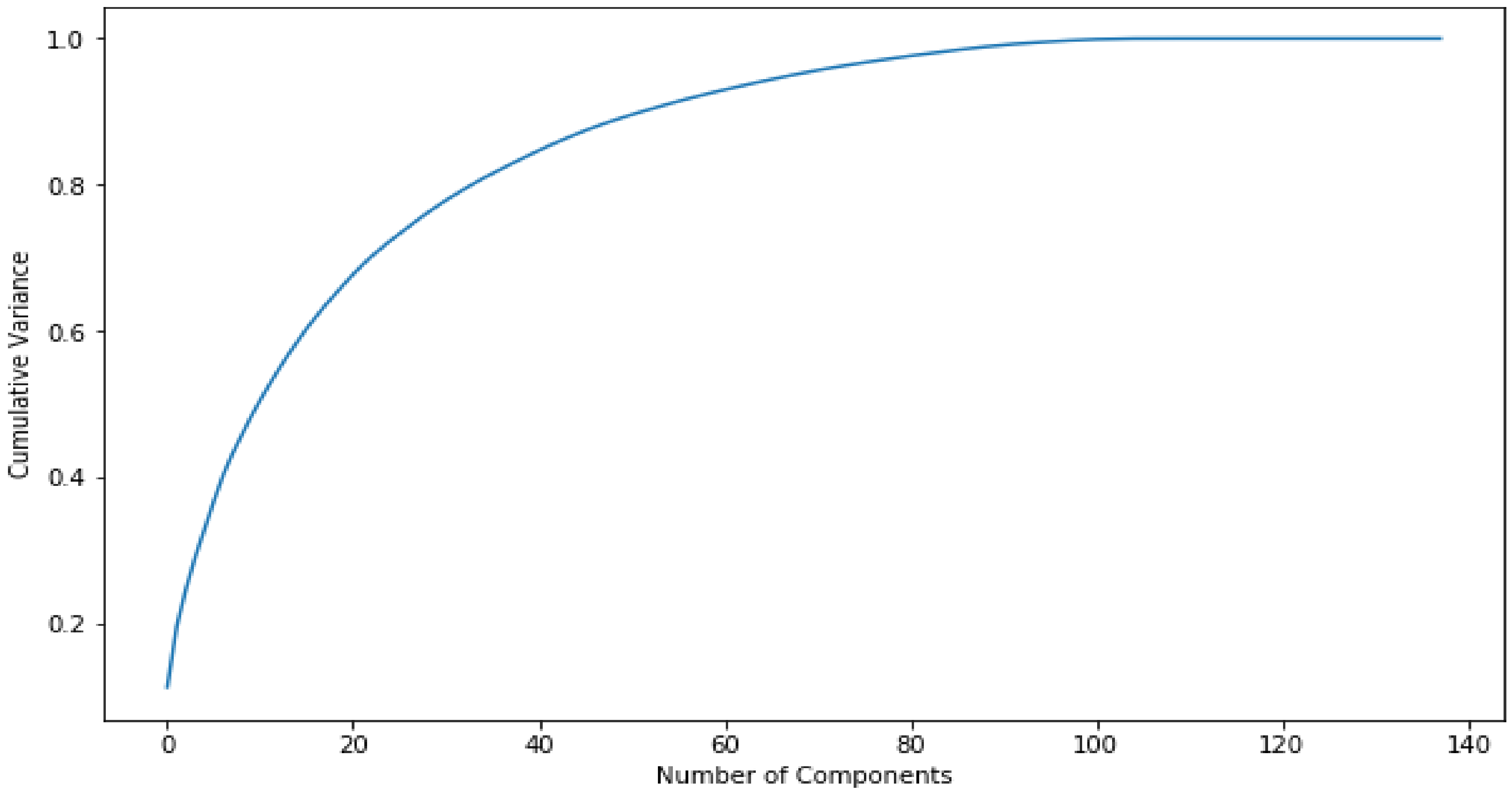


- ***Analysis of recharge amount and number of recharge in action month.***
- We can see from the above pattern that the recharge number and the recharge amount are mostly propotional. More the number of recharge, more the amount of the recharge.



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 Sesitivity has been decreased while evaluating the model on the test set. However, the accuracy and specificity is quite good in the test set.

Random forest with PCA

- ***Model summary***
- Train set
  - Accuracy = 0.84
  - Sensitivity = 0.88
  - Specificity = 0.80
- Test set
  - Accuracy = 0.80
  - Sensitivity = 0.75
  - Specificity = 0.80
- We can see from the model performance that the Sesitivity 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 acheiving 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 approx 81%. Also we have good accuracy of apporx 85%.

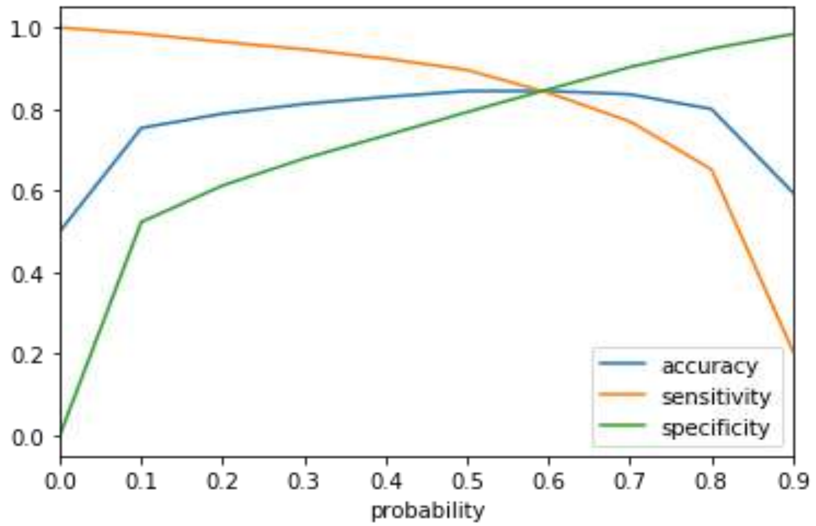
Model Without PCA

- **Logistic regression with No PCA.**

- ***Model analysis***
- We can see that there are few features have positive coefficients and few have negative.
- Many features have higher p-values and hence became insignificant in the model.
- ***Coarse tuning (Auto+Manual)***
- We'll first eliminate a few features using Recursive Feature Elimination (RFE), and once we have reached a small set of variables to work with, we can then use manual feature elimination (i.e. manually eliminating features based on observing the p-values and VIFs).

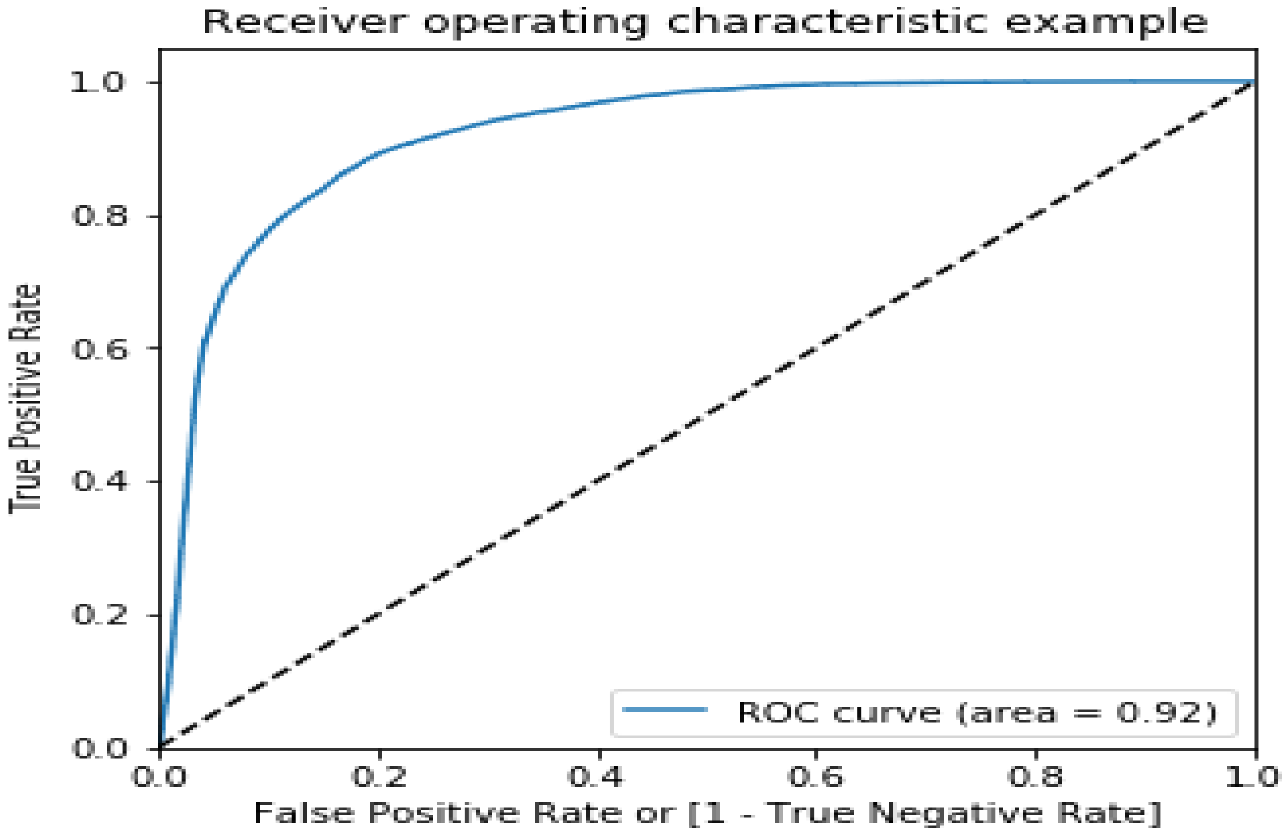
Feature Selection Using RFE

- **Analysis of the curve**
- Accuracy - Becomes stable around 0.6
- Sensitivity - Decreases with the increased probablity.
- Specificity - Increases with the increasing probablity.
- At point 0.6 where the three parameters cut each other, we can see that there is a balance bethween sensitivity and specificity with a good accuracy.
- Here we are intended to acheive 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 acheiving higher sensitivity, which is our main goal.



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

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



Testing the model on the test set

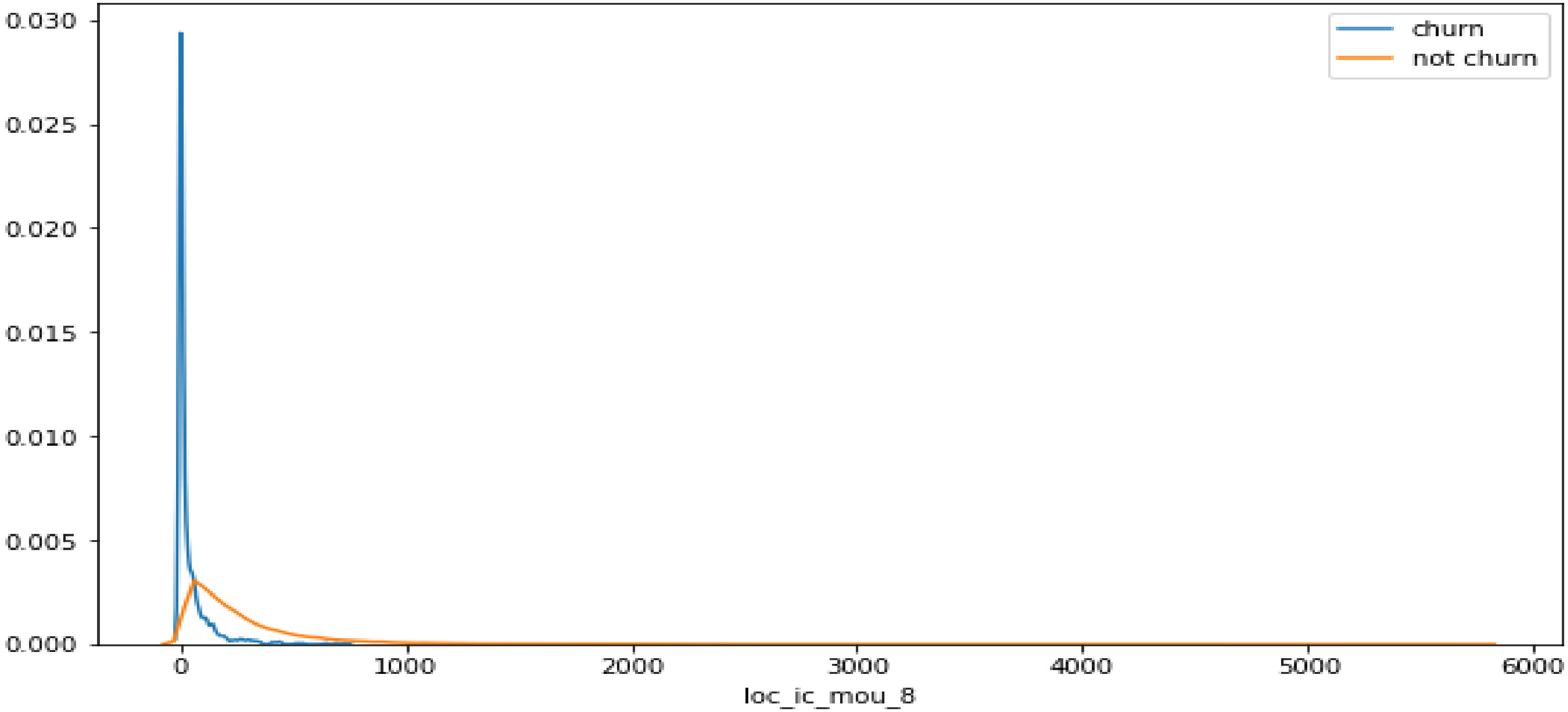
- Model summary
- Train set
  - Accuracy = 0.84
  - Sensitivity = 0.81
  - Specificity = 0.83
- Test set
  - Accuracy = 0.78
  - Sensitivity = 0.82
  - Specificity = 0.78
- Overall, the model is performing well in the test set, what it had learnt from the train set.

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.

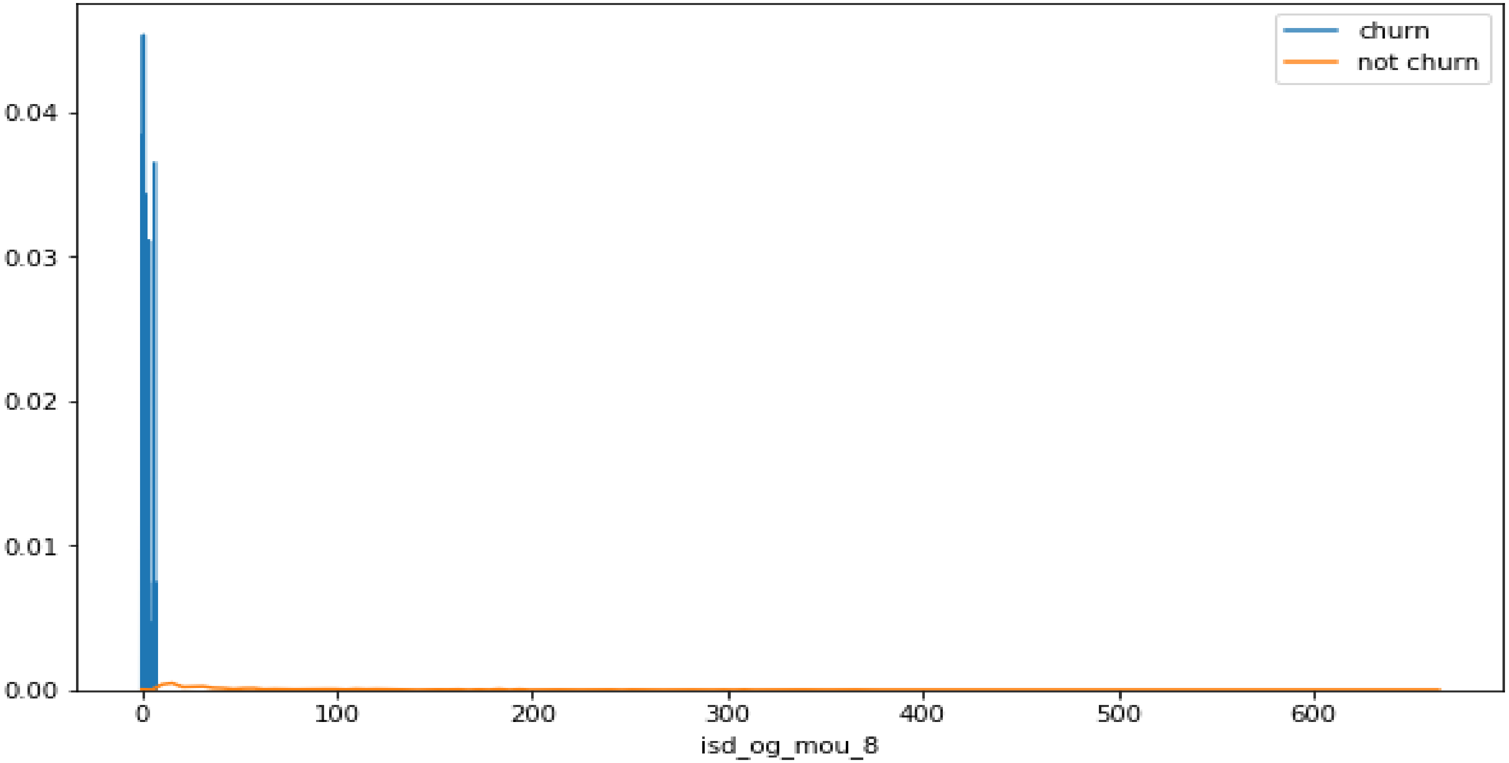
Plots of important predictors for churn and non churn customers

- 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.



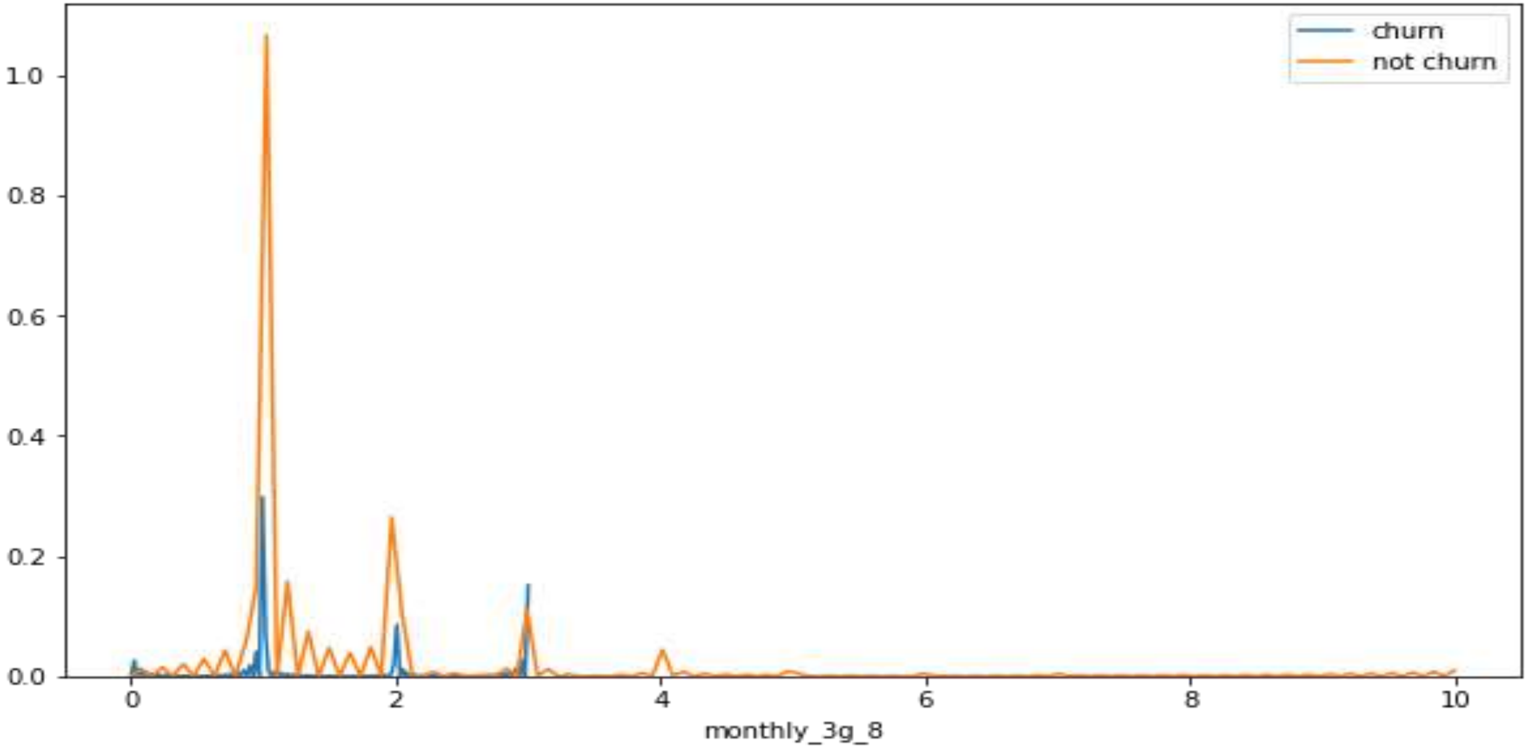
Plotting isd_og_mou_8 predictor for churn and not churn customers

- 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 onther 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 mothly 3g data for August for the churn customers are very much populated aroud 1, whereas of non churn customers it spreaded accross various numbers. Similarly, we can plot each variables, which have higher coefficients, churn distribution.



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.
- Cutomers, 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.
- Cutomers 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.

Thank You!