Telecom Churn Case Study

Matin Shaikh

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.

High-value churn

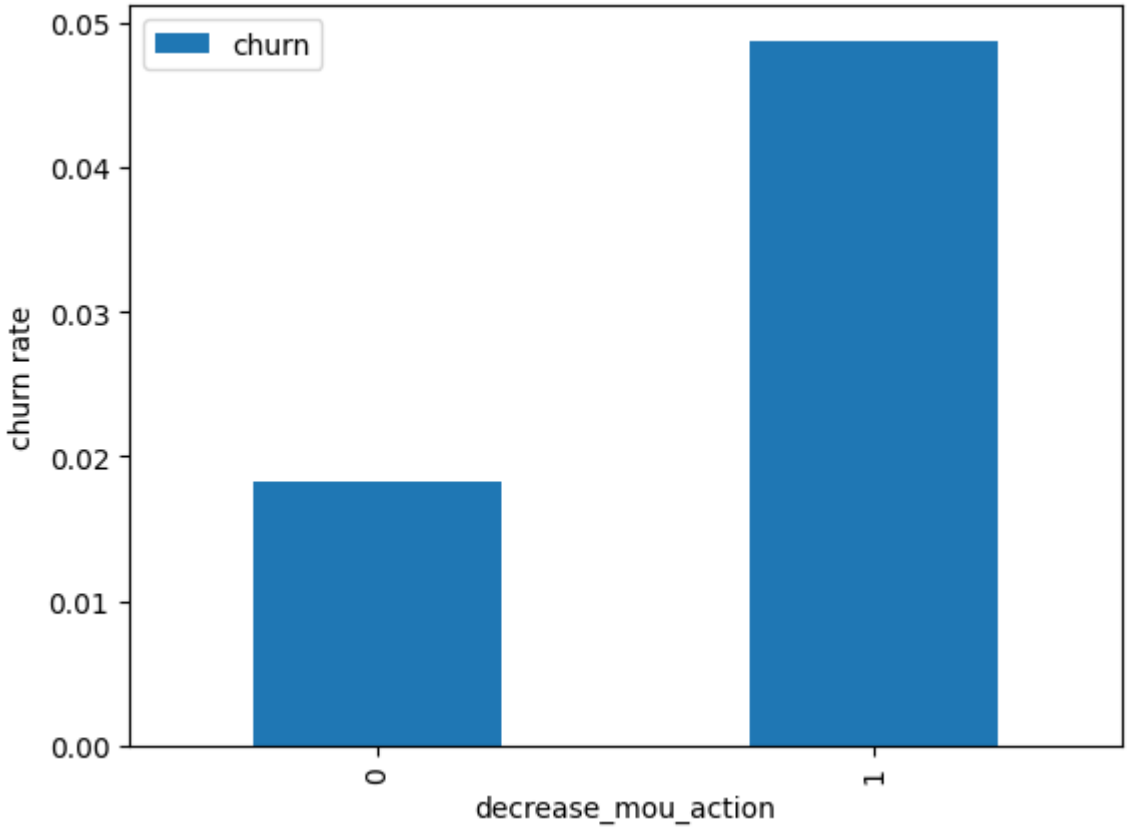
In the Indian and Southeast Asian markets, approximately 80% of revenue comes from the top 20% of customers (called high-value customers). Thus, if we can reduce the churn of high-value customers, we will be able to reduce significant revenue leakage.

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.

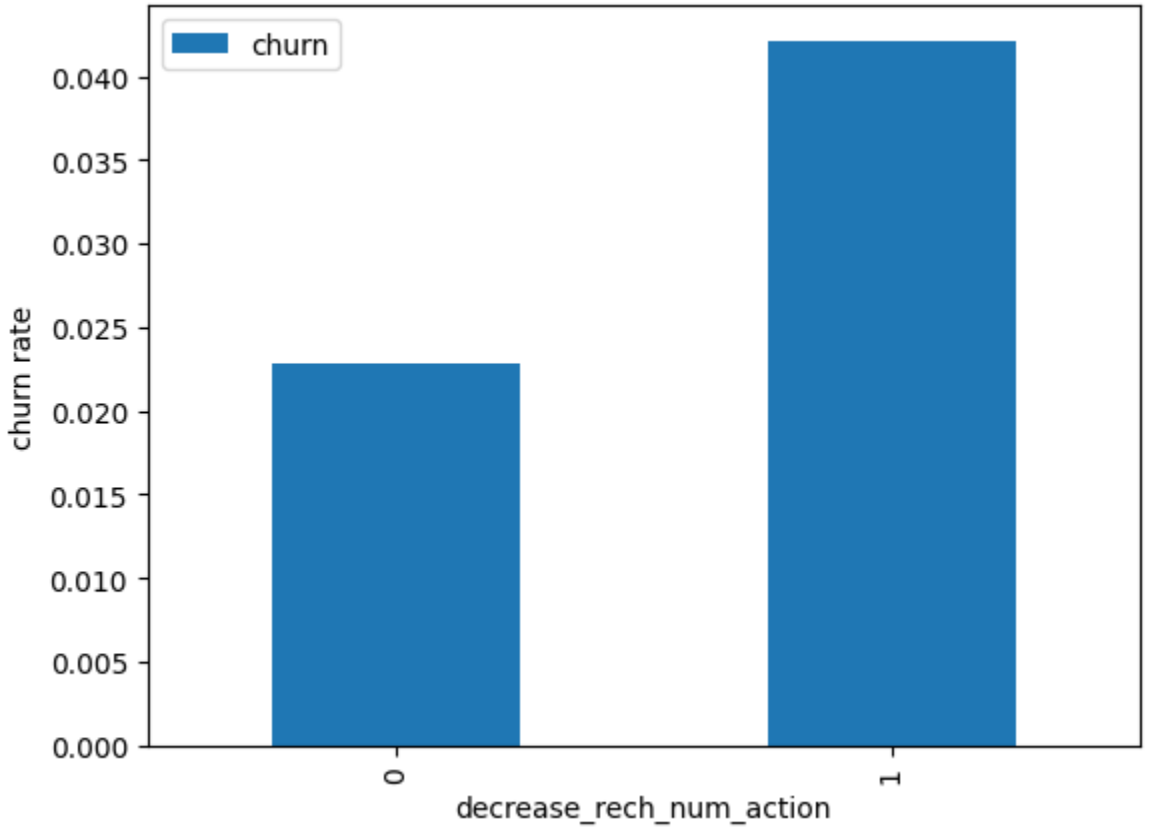
Business objective

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 behaviour during churn will be helpful.

Exploratory Data Analysis (EDA) - Univariate Analysis

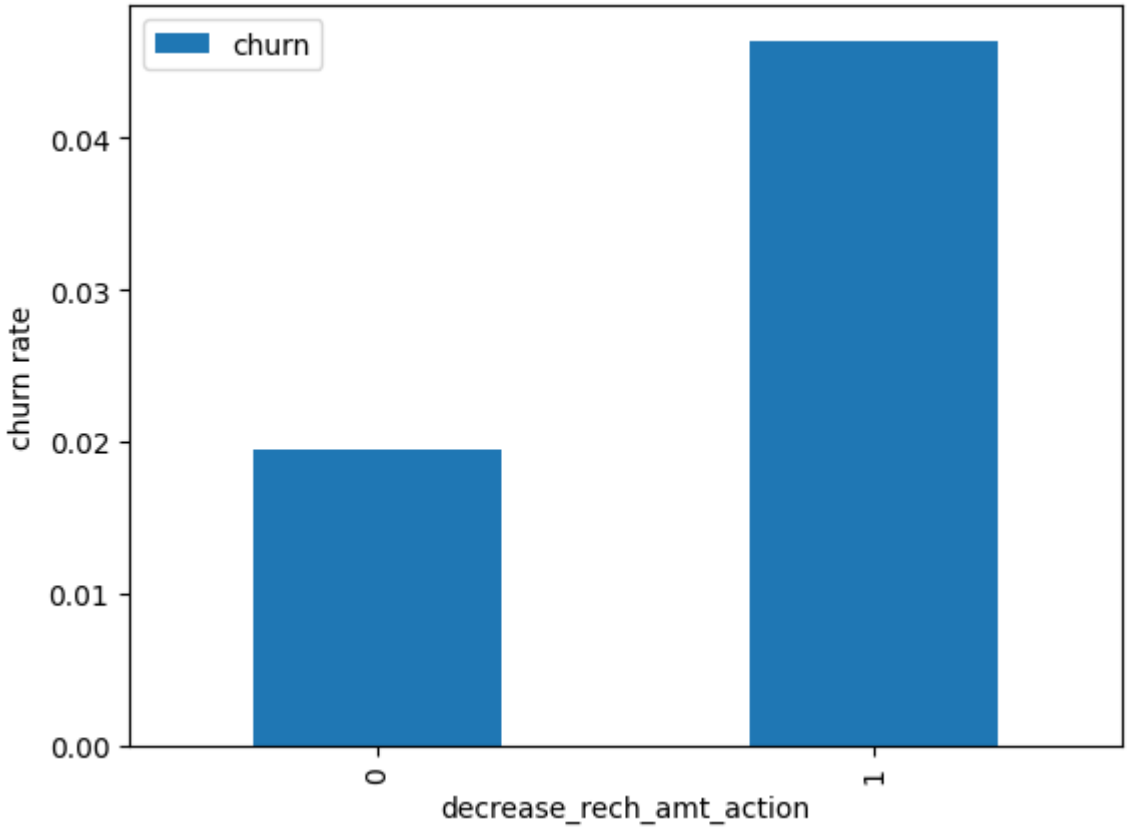


We can see that the churn rate is more for the customers, whose minutes of usage(MOU) decreased in the action phase than the good phase.

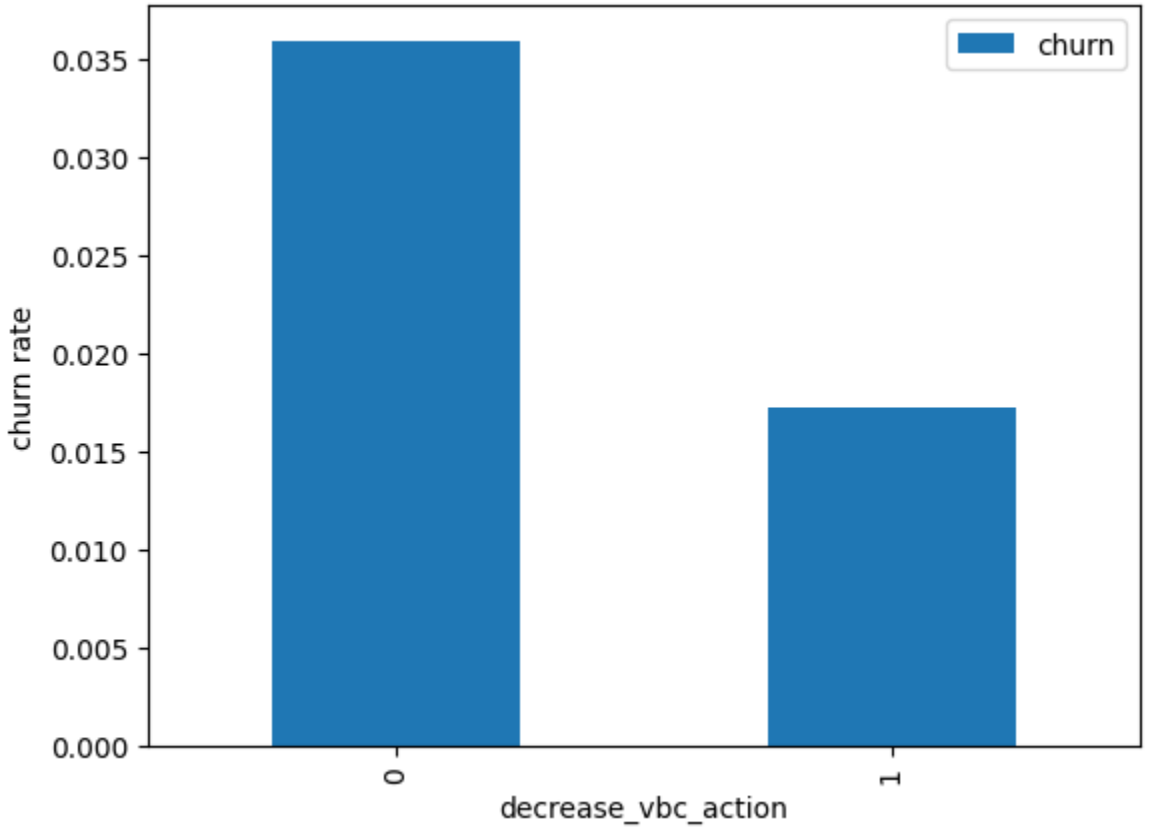


As expected, the churn rate is more for the customers, whose number of recharge in the action phase is lesser than the number in good phase.

Exploratory Data Analysis (EDA) - Univariate Analysis

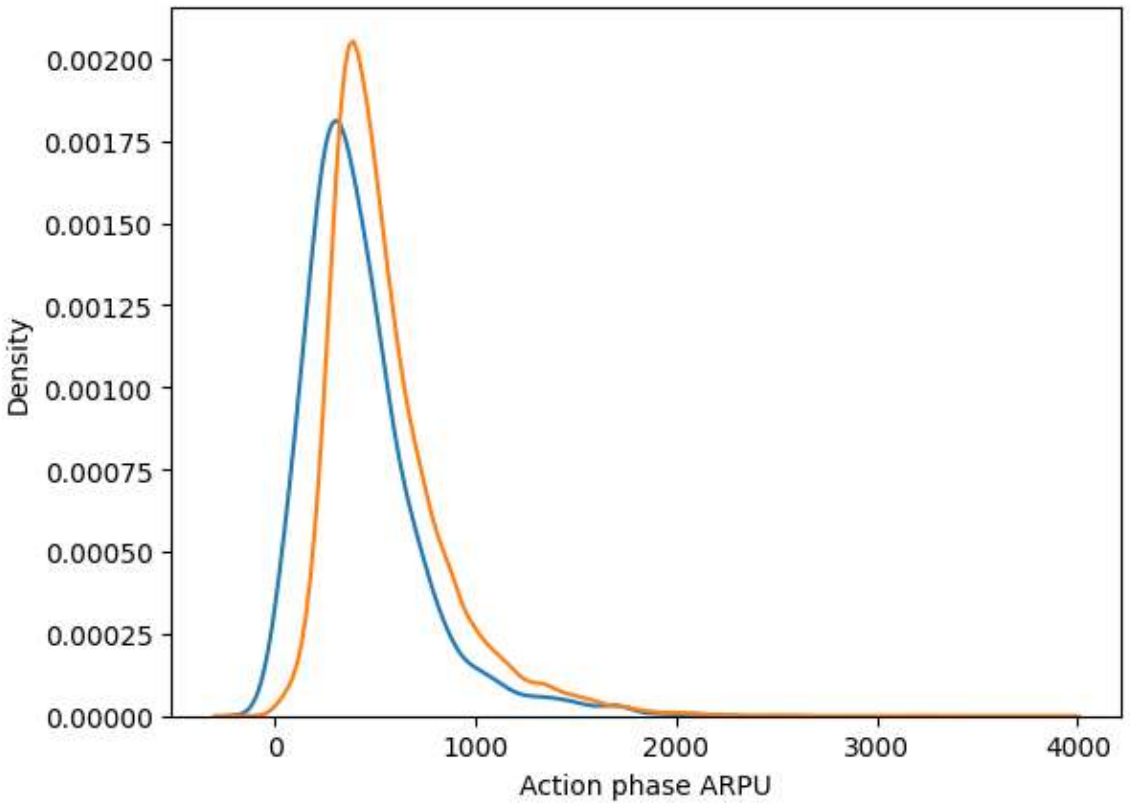


The churn rate is more for the customers, whose amount of recharge in the action phase is lesser than the amount in good phase.

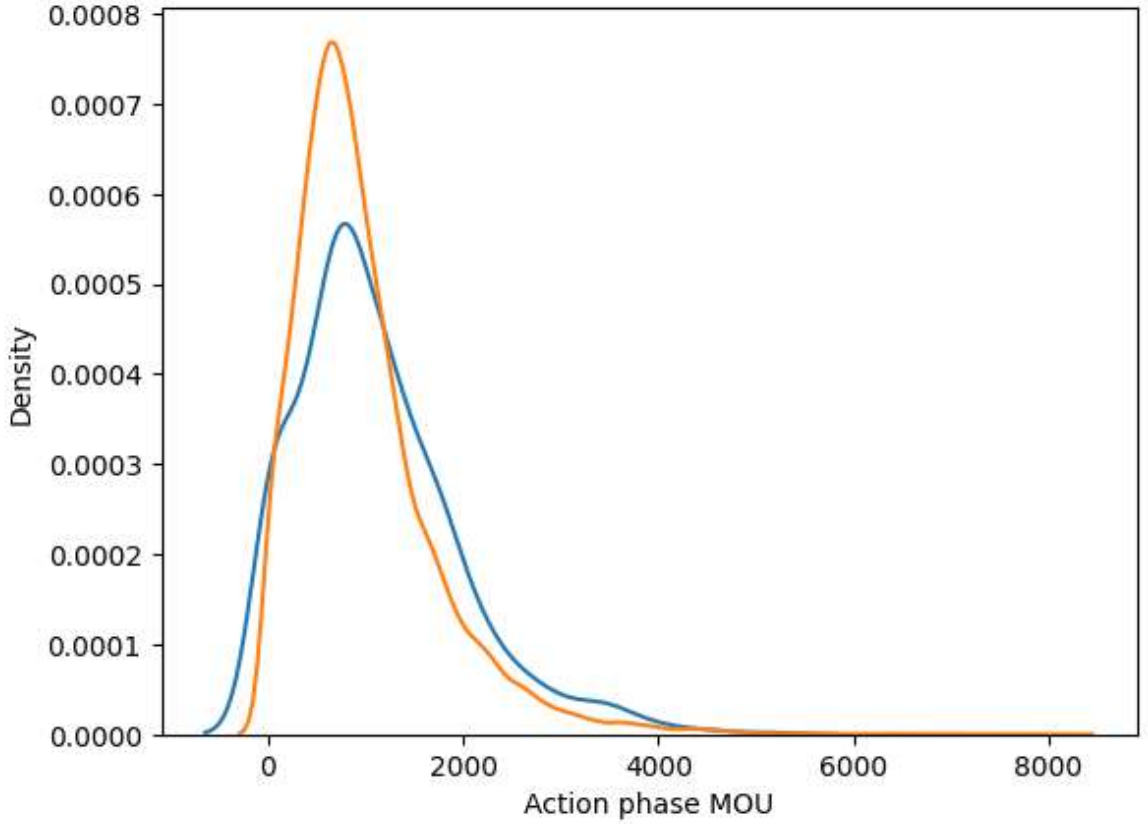


The churn rate is more for the customers, whose volume based cost in action month is increased. That means the customers do not do the monthly recharge more when they are in the action phase.

Exploratory Data Analysis (EDA) - Univariate Analysis

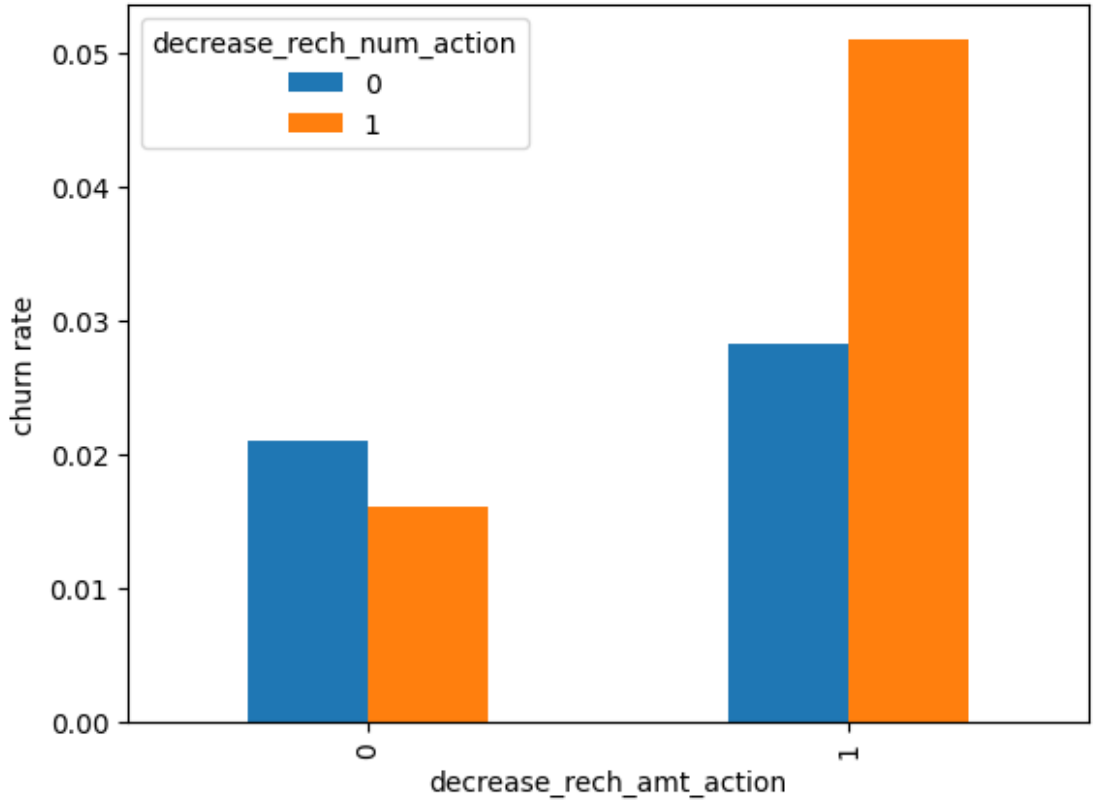


Average revenue per user (ARPU) for the churned customers is mostly densed on the 0 to 900. The higher ARPU customers are less likely to be churned.
ARPU for the not churned customers is mostly densed on the 0 to 1000.

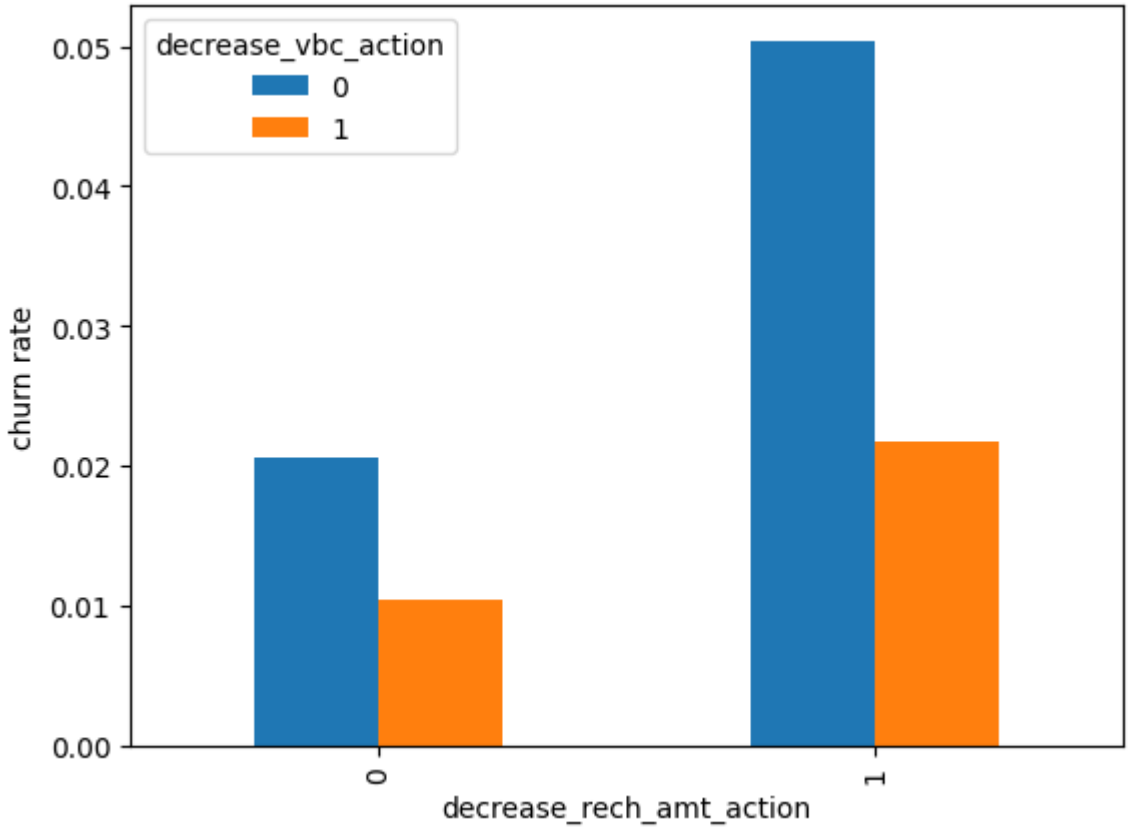


Minutes of usage(MOU) of the churn customers is mostly populated on the 0 to 2500 range. Higher the MOU, lesser the churn probability.

Exploratory Data Analysis (EDA) - Bivariate Analysis

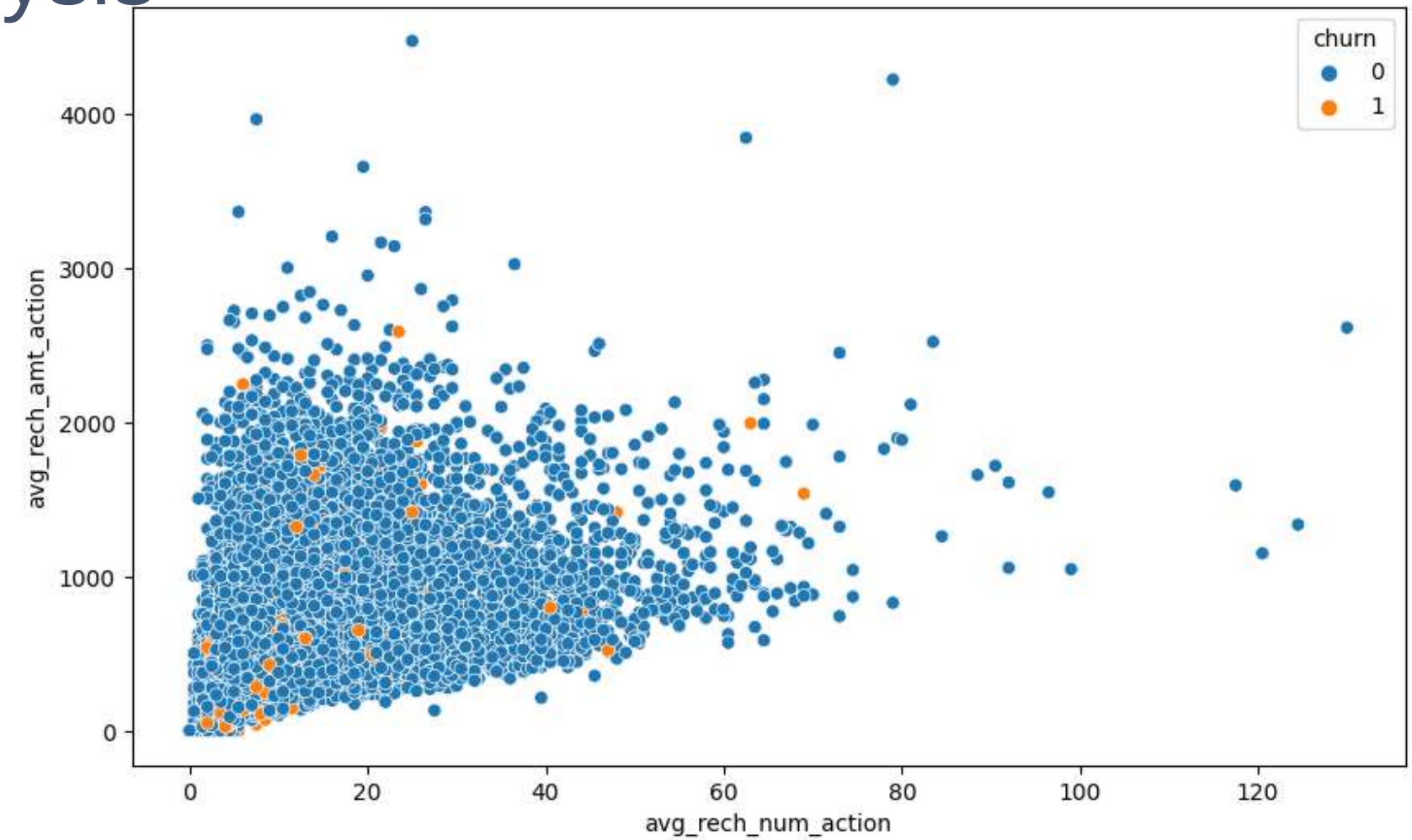


The churn rate is more for the customers, whose recharge amount as well as number of recharge have decreased in the action phase than the good phase.



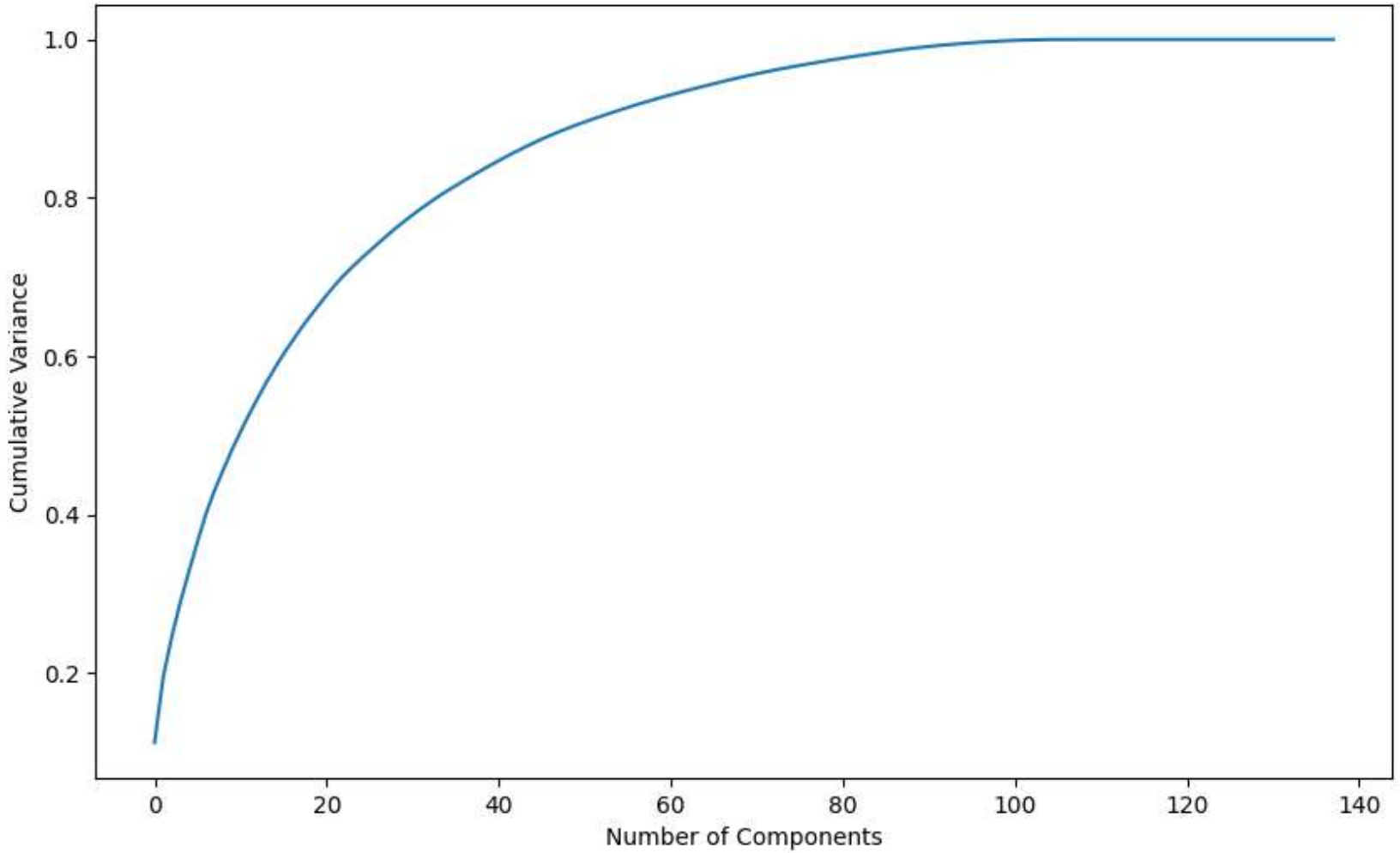
Here, also we can see that the churn rate is more for the customers, whose recharge amount is decreased along with the volume based cost is increased in the action month.

Exploratory Data Analysis (EDA) - Bivariate Analysis



The above pattern shows that the recharge number and the recharge amount are mostly proportional. More the number of recharge, more the amount of the recharge.

Model with Principal Component Analysis (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 = 0.87

Sensitivity = 0.90

Specificity = 0.84

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.

Decision tree with PCA

Model summary:

Train set:

- Accuracy = 0.90
- Sensitivity = 0.92
- Specificity = 0.88

Test set

- Accuracy = 0.86
- Sensitivity = 0.70
- Specificity = 0.87

The model performance for Sensitivity has decreased 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 = 0.85
- Sensitivity = 0.89
- Specificity = 0.81

Test set

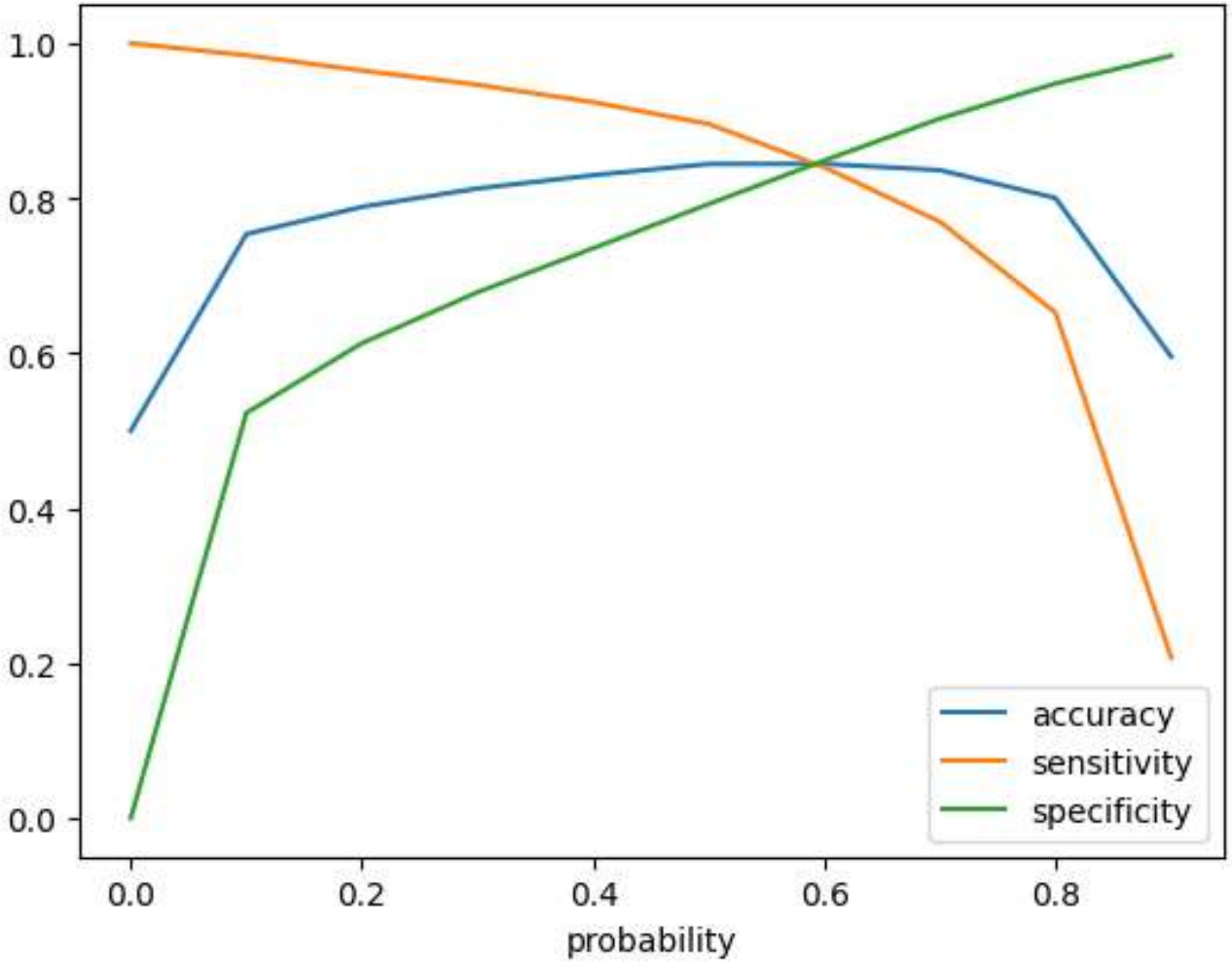
- Accuracy = 0.80
- Sensitivity = 0.76
- Specificity = 0.80

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.

Conclusion On Model with PCA

After trying several model’s, we can see that for achieving the best sensitivity, which was our goal, the classic Logistic regression models preforms well. The sensitivity was approx. 81%. Also, we have good accuracy of approx. 83%.

Logistic regression with No PCA



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.

Logistic regression with No PCA

Model summary:

Train set:

- Accuracy = 0.84
- Sensitivity = 0.89
- Specificity = 0.79

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.

Conclusion On Model 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 expliains the important predictor variables as well as the significance of each variable. The model also hels 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.

Business Recommendations

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

Top predictors

Below 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 probablity.

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.

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 charge in 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 with more monthly 3G recharge in August 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.