Domain Oriented Case Study

- Telecom Churn

– Akshara J

TABLE OF CONTENTS

- Overview
- Understanding & Defining Churn
- Understanding Customer Behavior During Churn
- Data preparation
- Exploratory data analysis
- Model Summary conclusion with PCA for Logistic regression, Support Vector Machine(SVM) with PCA, Decision Tree, Random Forest
- Logistic regression with No PCA
- Final model summary with top predictors and business recommendation

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 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.

UNDERSTANDING & 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 we 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).
- There are various ways to define churn, such as:
    - Revenue-based churn
    - Usage-based churn

CUSTOMER BEHAVIOR DURING CHURN

In churn prediction, we assume that there are three phases of customer lifecycle:

- **The ‘good’ phase:** In this phase, the customer is happy with the service and behaves as usual.
- **The ‘action’ phase:** The customer experience starts to sore in this phase, for e.g. he/she gets a compelling offer from a competitor, faces unjust charges, becomes unhappy with service quality etc. In this phase, the customer usually shows different behaviour than the ‘good’ months. Also, it is crucial to identify high-churn-risk customers in this phase, since some corrective actions can be taken at this point (such as matching the competitor’s offer/improving the service quality etc.)
- **The ‘churn’ phase:** In this phase, the customer is said to have churned. We define churn based on this phase. Also, it is important to note that at the time of prediction (i.e. the action months), this data is not available to us for prediction. Thus, after tagging churn as 1/0 based on this phase, we discard all data corresponding to this phase.
- In this case, since we are working over a four-month window, the first two months are the ‘good’ phase, the third month is the ‘action’ phase, while the fourth month is the ‘churn’ phase.

DATA PREPARATION

The following data preparation steps are crucial for this problem:

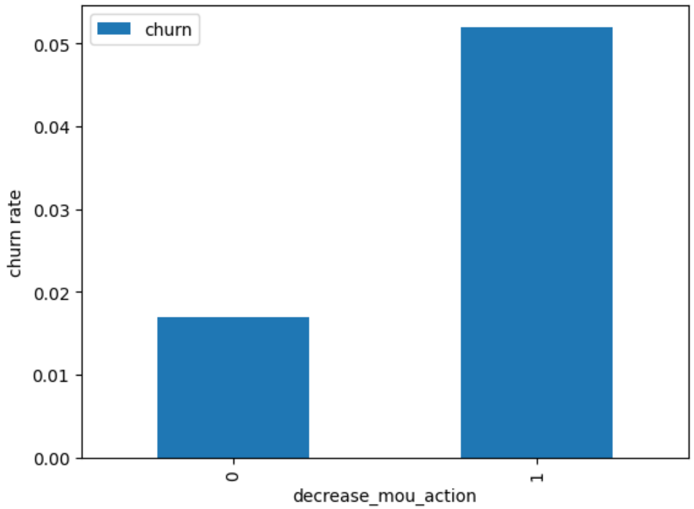
1. **Filter high-value customers:**
    1. Define high-value customers as follows: Those who have recharged with an amount more than or equal to X, where X is the 70th percentile of the average recharge amount in the first two months (the good phase).
    2. X value defined is 369.5

2. 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 you need to use to tag churners are:

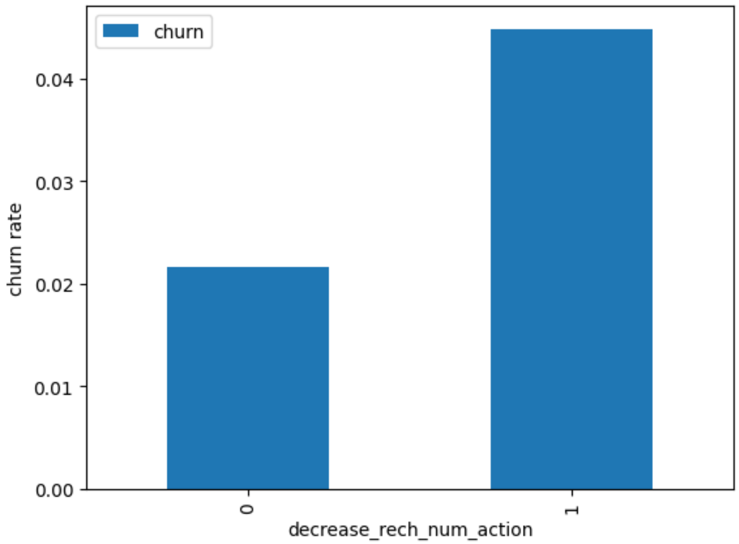
total_ic_mou_9
total_og_mou_9
vol_2g_mb_9
vol_3g_mb_9

After tagging churners, remove all the attributes corresponding to the churn phase (all attributes having ‘ _9’, etc. in their names).

EXPLORATORY DATA ANALYSIS

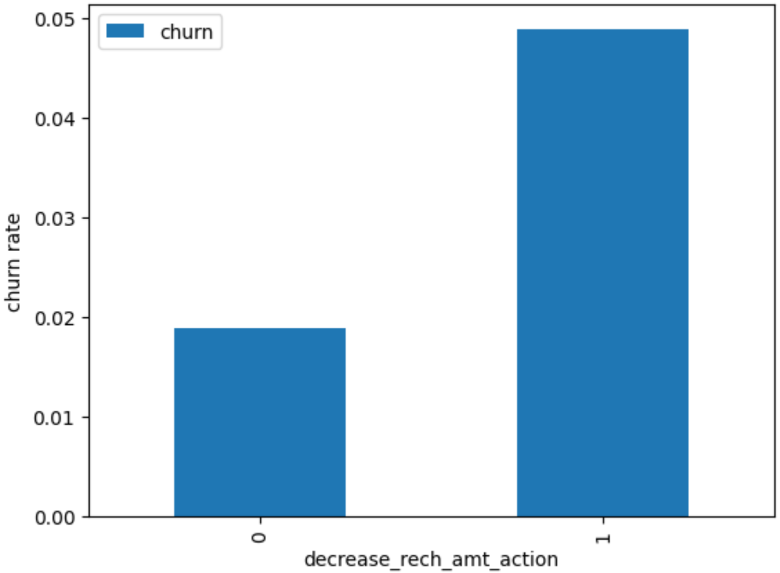


From the above table, the churn rate is more for the customers, whose minutes of usage(mou) is decreased in the action phase when compared to the good phase

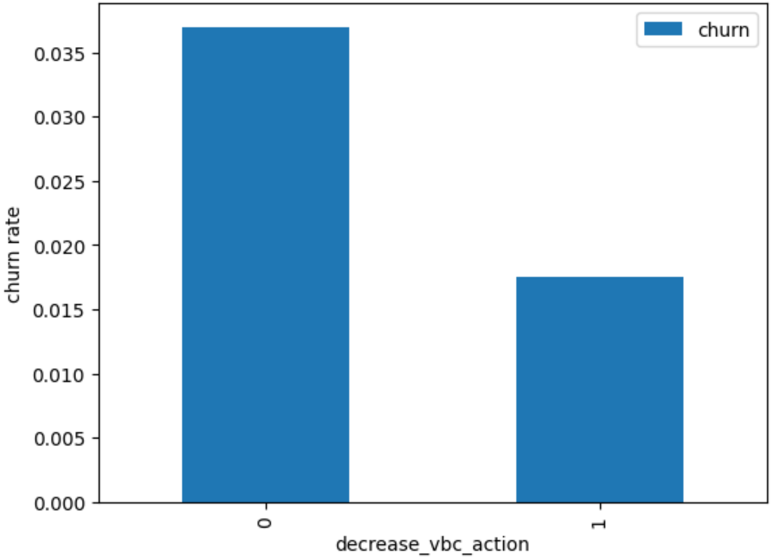


As expected, the churn rate is more for the customers, whose number of recharge in the action phase is lesser than the number in good phase.

EXPLORATORY DATA ANALYSIS

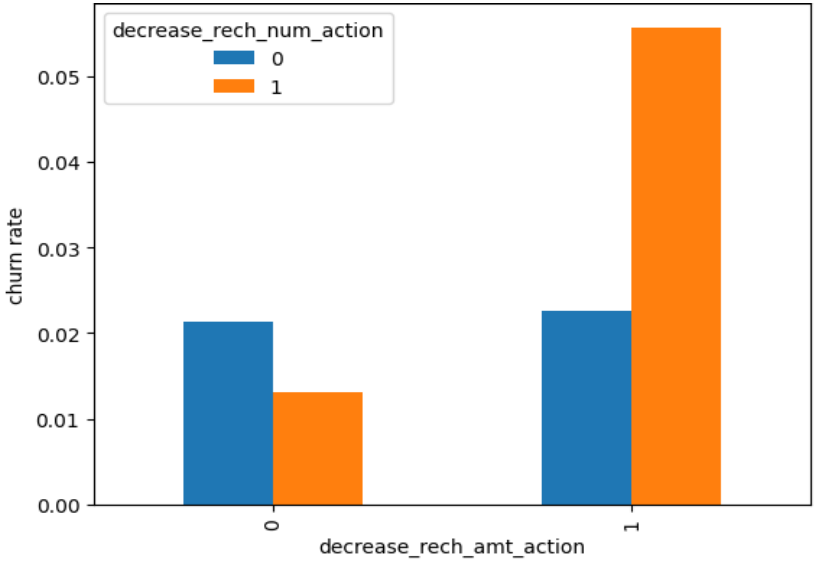


Here also we see the same behaviour. The churn rate is more for the customers, whose amount of recharge in the action phase is lesser than the amount in good phase.

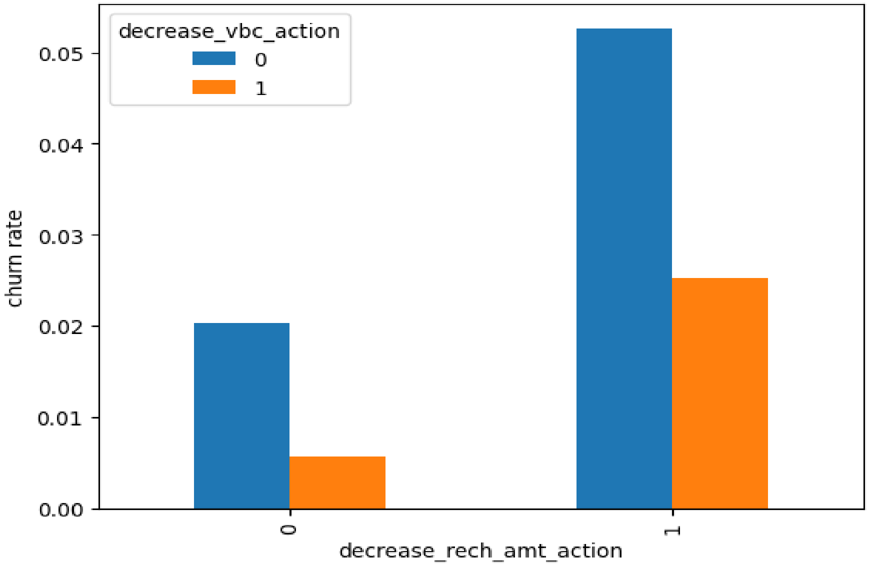


Here we see the expected result. The churn rate is more for the customers, whose volume based cost in action month is increased. That means the customers do not do the monthly recharge more when they are in the action phase.

EXPLORATORY DATA 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.



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.

LOGISTIC REGRESSION	SUPPORT VECTOR MACHINE(SVM) WITH PCA

Model summary
Train set
Accuracy:- 0.87
Sensitivity:- 0.89
Specificity:- 0.84
Test set
Accuracy:- 0.84
Sensitivity:- 0.82
Specificity:- 0.84
Overall, the model is performing well in the test set, what it had learnt from the train set.
	Model summary
	Train set
	Accuracy = 0.89
	Sensitivity = 0.91
	Specificity = 0.87
	Test set
	Accuracy = 0.86
	Sensitivity = 0.78
	Specificity = 0.86

DECISION TREE WITH PCA	RANDOM FOREST WITH PCA

Model summary
Train set
Accuracy = 0.91
Sensitivity = 0.92
Specificity = 0.89
Test set
Accuracy = 0.86
Sensitivity = 0.72
Specificity = 0.86
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.
	Model summary
	Train set
	Accuracy = 0.88
	Sensitivity = 0.89
	Specificity = 0.86
	Test set
	Accuracy = 0.84
	Sensitivity = 0.73
	Specificity = 0.85
	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 approx. 90%. Also we have good accuracy of approx. 88%.

LOGISTIC REGRESSION WITH NO PCA

Model analysis

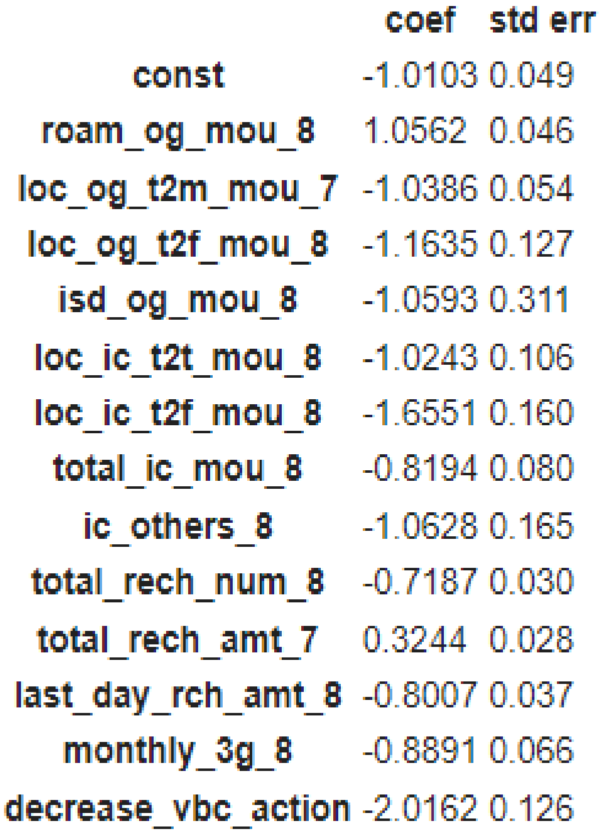
1.We can see that there are few features have positive coefficients and few have negative.
2.Many features have higher p-values and hence became insignificant in the model.

Coarse tuning (Auto+Manual)

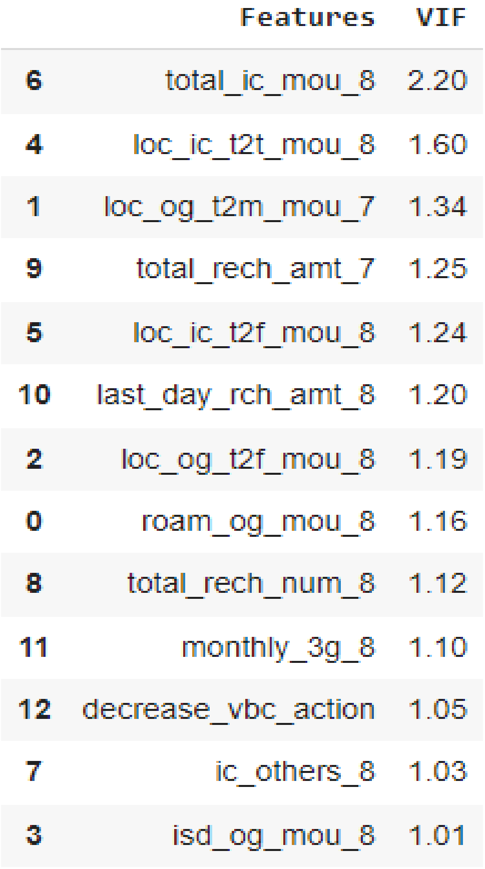
We'll first eliminate a few features using Recursive Feature Elimination (RFE), and once we have reached a small set of variables to work with, we can then use manual feature elimination (i.e. manually eliminating features based on observing the p-values and VIFs).

FINAL MODEL III

From the model summary and the VIF list we can see that all the variables are significant and there is no multicollinearity among the variables.
Hence, we can conclude that *Model-3 log_no_pca_3 will be the final model*.

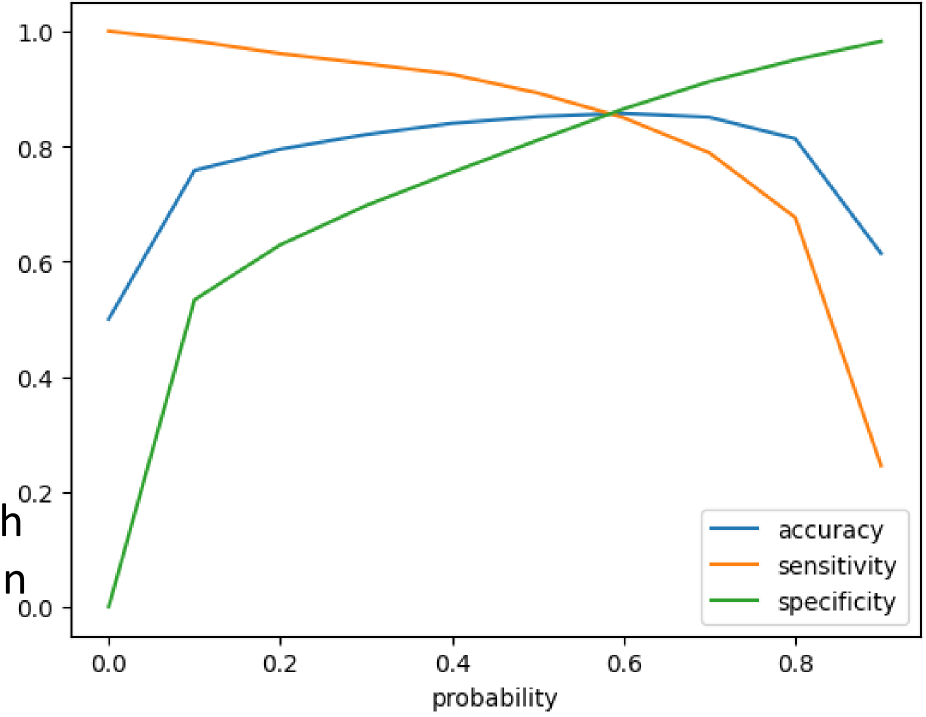


	coef	std err
const	-1.0103	0.049
roam_og_mou_8	1.0562	0.046
loc_og_t2m_mou_7	-1.0386	0.054
loc_og_t2f_mou_8	-1.1635	0.127
isd_og_mou_8	-1.0593	0.311
loc_ic_t2t_mou_8	-1.0243	0.106
loc_ic_t2f_mou_8	-1.6551	0.160
total_ic_mou_8	-0.8194	0.080
ic_others_8	-1.0628	0.165
total_rech_num_8	-0.7187	0.030
total_rech_amt_7	0.3244	0.028
last_day_rch_amt_8	-0.8007	0.037
monthly_3g_8	-0.8891	0.066
decrease_vbc_action	-2.0162	0.126



	Features	VIF
6	total_ic_mou_8	2.20
4	loc_ic_t2t_mou_8	1.60
1	loc_og_t2m_mou_7	1.34
9	total_rech_amt_7	1.25
5	loc_ic_t2f_mou_8	1.24
10	last_day_rch_amt_8	1.20
2	loc_og_t2f_mou_8	1.19
0	roam_og_mou_8	1.16
8	total_rech_num_8	1.12
11	monthly_3g_8	1.10
12	decrease_vbc_action	1.05
7	ic_others_8	1.03
3	isd_og_mou_8	1.01

OPTIMAL PROBABILITY CUTOFF POINT



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.

MODEL SUMMARY

Train set

Accuracy = 0.85
Sensitivity = 0.89
Specificity = 0.81

Test set

Accuracy = 0.82
Sensitivity = 0.83
Specificity = 0.82

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

Final conclusion with no PCA

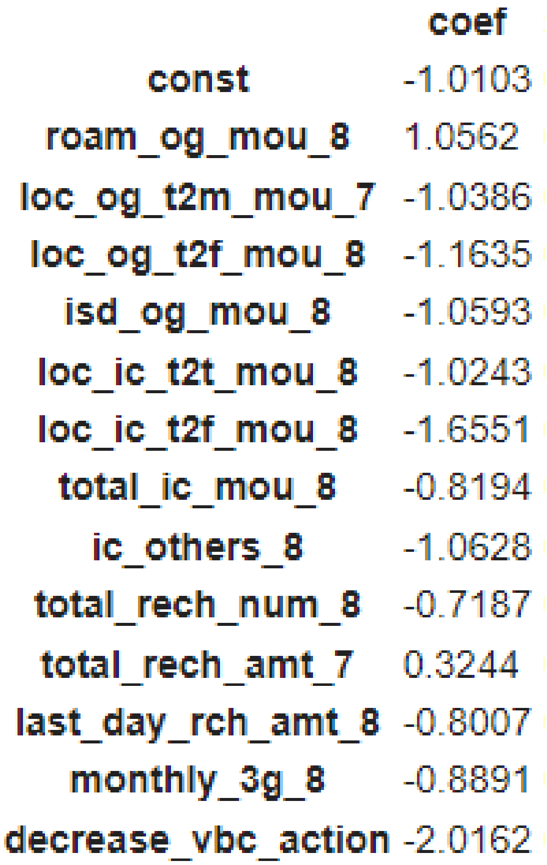
- We can see that the logistic model with no PCA has good sensitivity and accuracy, which are comparable to the models with PCA.
- So, we can go for the more simplistic model such as logistic regression with PCA as it explains the important predictor variables as well as the significance of each variable.
- 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.

TOP PREDICTORS

These 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.

For example,
If the total incoming minutes of usage (total_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.



	coef
const	-1.0103
roam_og_mou_8	1.0562
loc_og_t2m_mou_7	-1.0386
loc_og_t2f_mou_8	-1.1635
isd_og_mou_8	-1.0593
loc_ic_t2t_mou_8	-1.0243
loc_ic_t2f_mou_8	-1.6551
total_ic_mou_8	-0.8194
ic_others_8	-1.0628
total_rech_num_8	-0.7187
total_rech_amt_7	0.3244
last_day_rch_amt_8	-0.8007
monthly_3g_8	-0.8891
decrease_vbc_action	-2.0162

RECOMMENDATION

1. 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).

2. Target the customers, whose outgoing others charge in July and incoming others on August are less.

3. 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.

4. Customers having decreasing local incoming minutes of usage for operators T to fixed lines of T for the month of August are more likely to churn.

5. Customers decreasing monthly 3g usage for August are most probable to churn.

6. Customers having decreasing incoming minutes of usage for operators T to fixed lines of T for August are more likely to churn.

7. roam_og_mou_8 variables have positive coefficients (2.1247). That means for the customers, whose roaming outgoing minutes of usage is increasing are more likely to churn.