Telecom Churn Case Study

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.

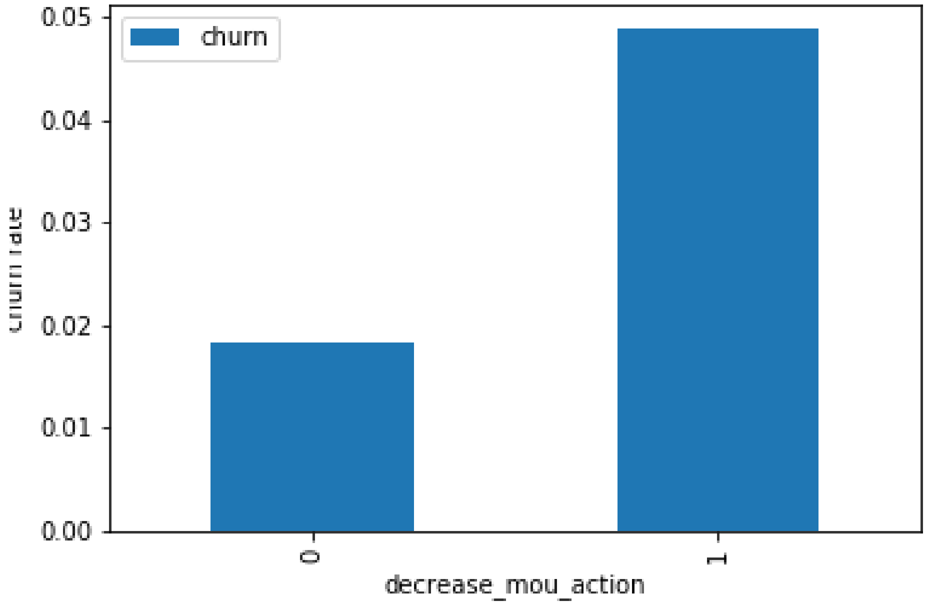
For many incumbent operators, retaining high profitable customers is the number one business goal.

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.

EDA

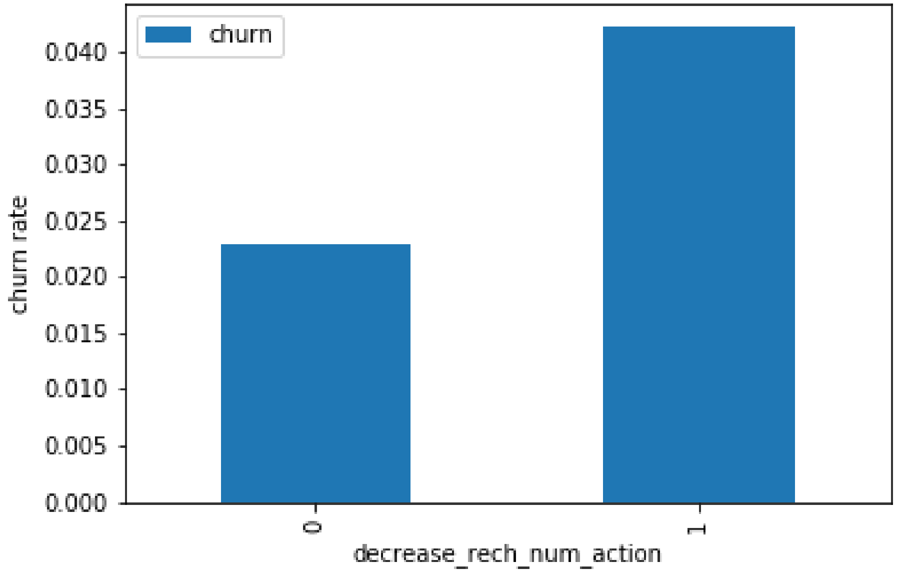
Univariate analysis



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.

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



Analysis

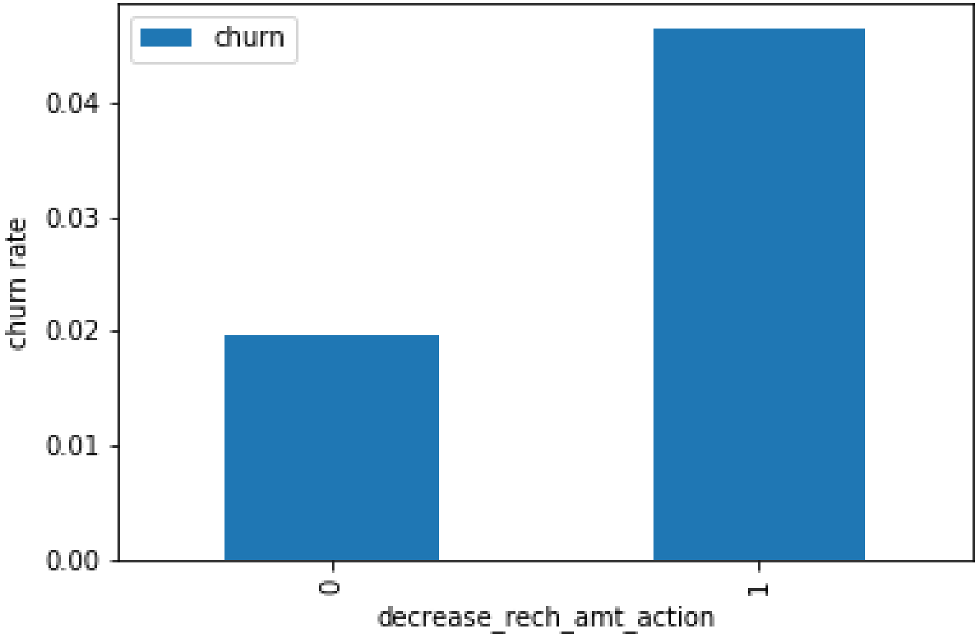
As expected, the churn rate is more for the customers, whose number of recharge in the action phase is lesser than the number in good phase.

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

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.

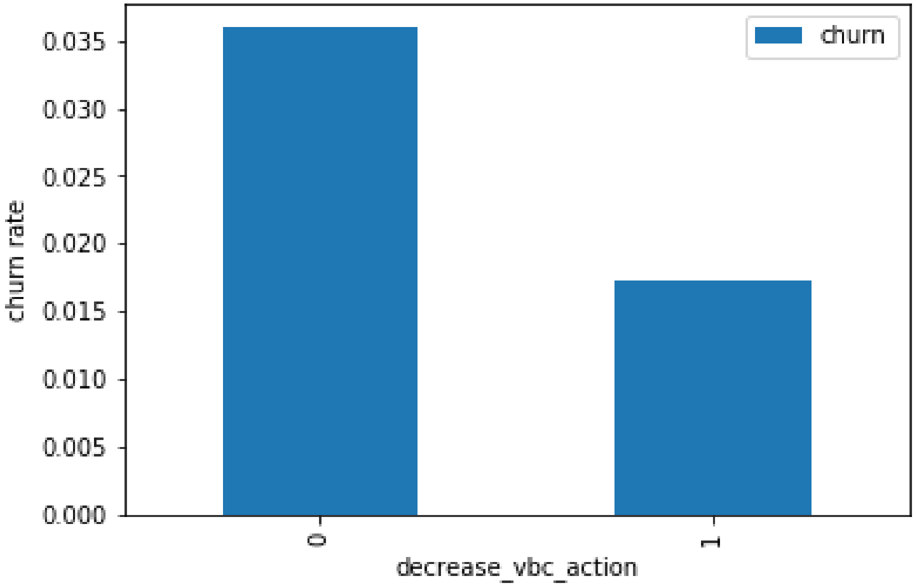
Churn rate on the basis whether the customer decreased her/his volume based cost in action month



Analysis

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

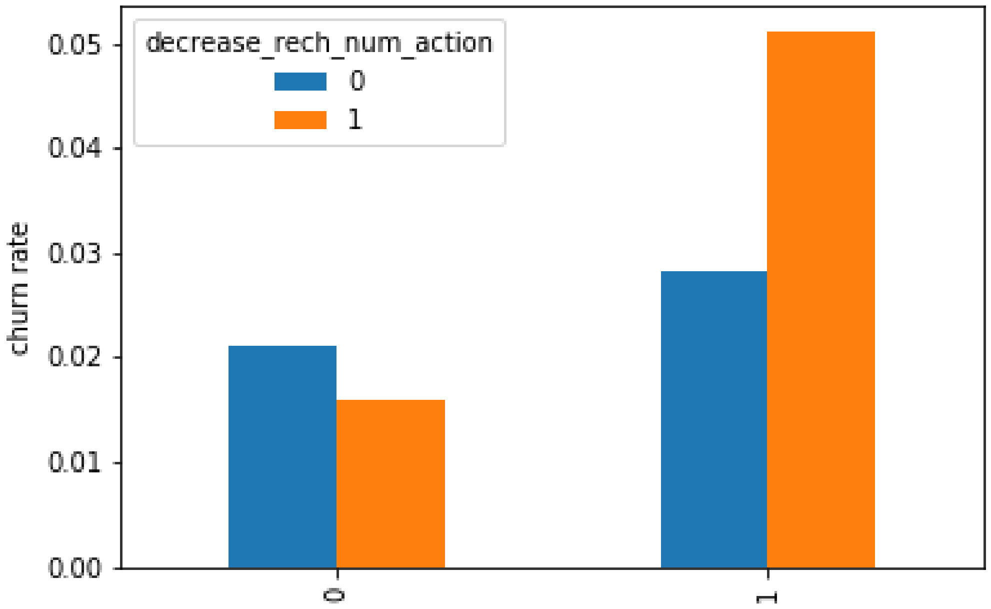


Bivariate analysis

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.

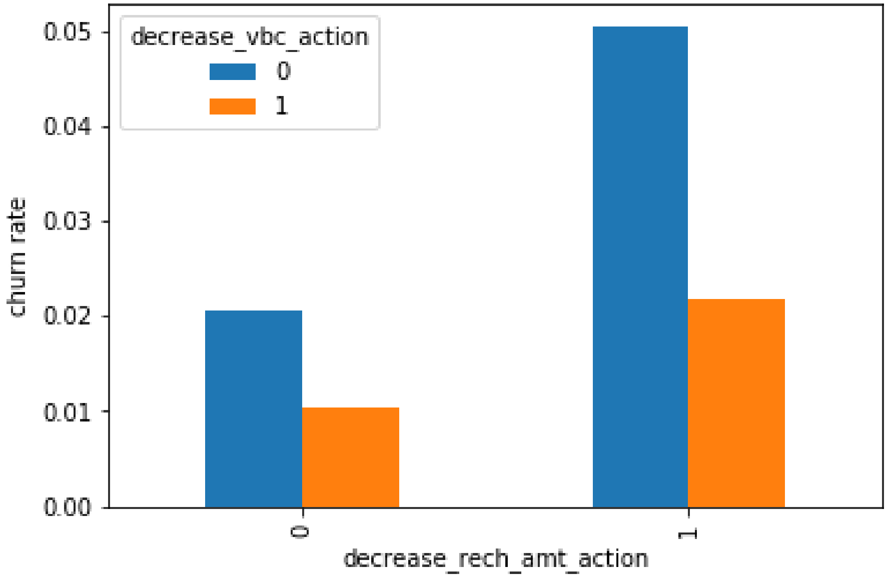
Analysis of churn rate by the decreasing recharge amount and volume based cost in the action phase



Analysis

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



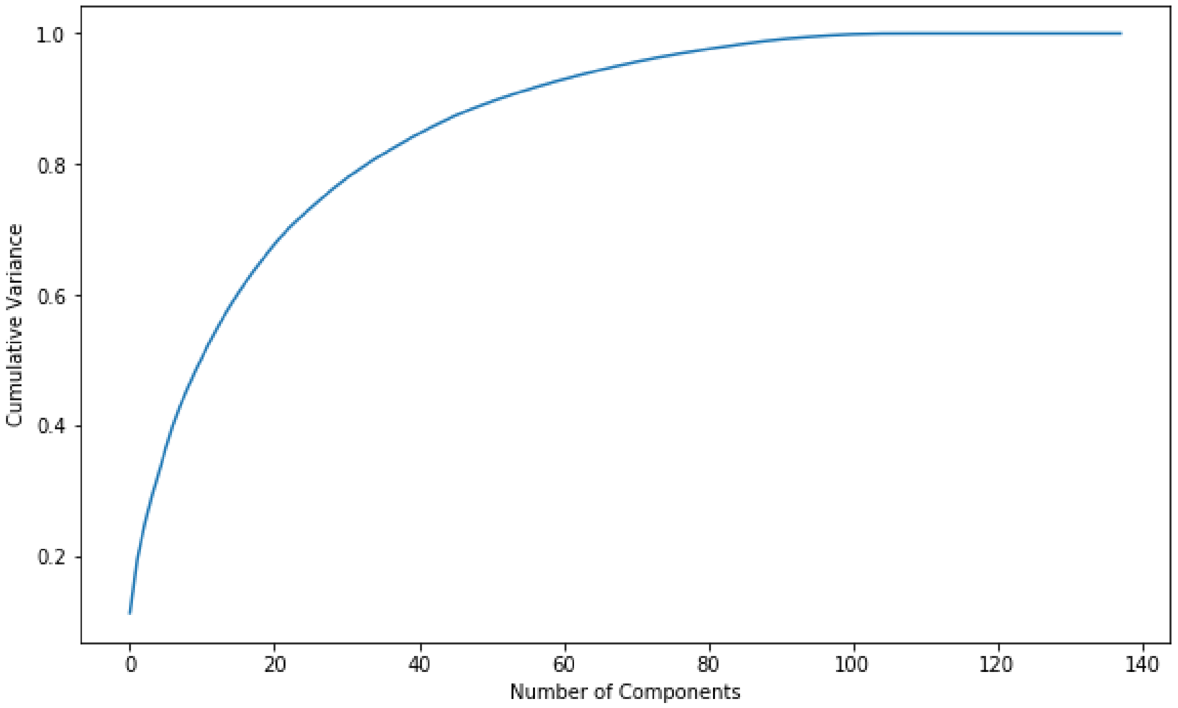
Model with PCA

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

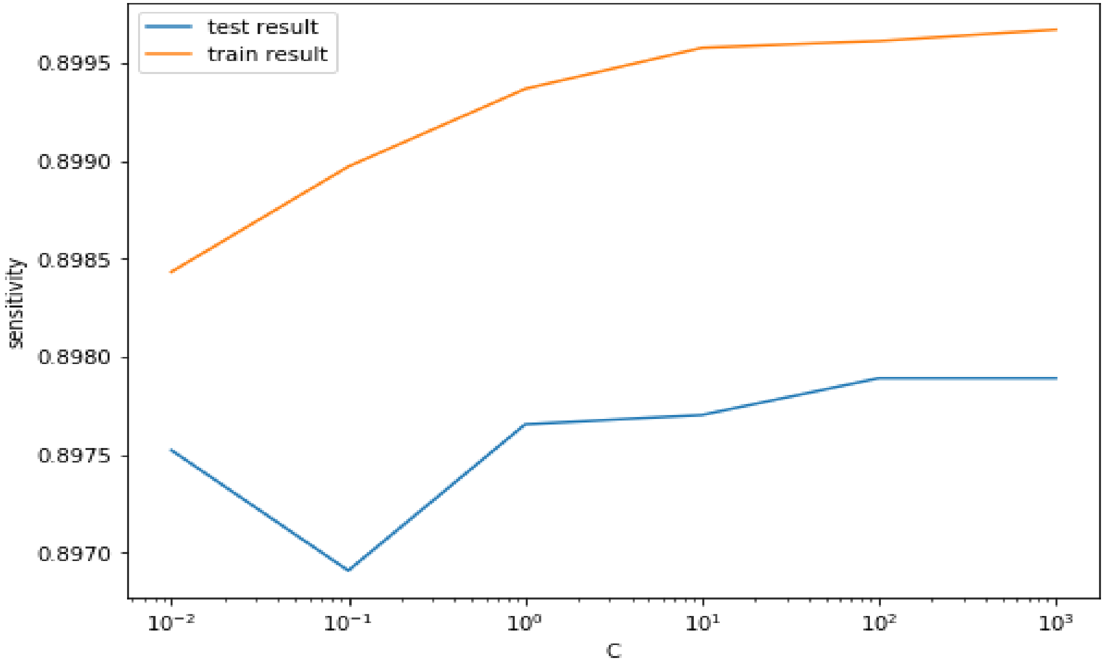
Emphasize Sensitivity/Recall than Accuracy

We are more focused on higher Sensitivity/Recall score than the accuracy.

Beacuse we need to care more about churn cases than the not churn cases. The main goal is to reatin the customers, who have the possiblity to churn. There should not be a problem, if we consider few not churn customers as churn customers and provide them some incentives for retaining them. Hence, the sensitivity score is more important here.



Logistic regression with PCA
Tuning hyperparameter C:- is the the inverse of regularization strength in Logistic Regression. Higher values of C correspond to less regularization.

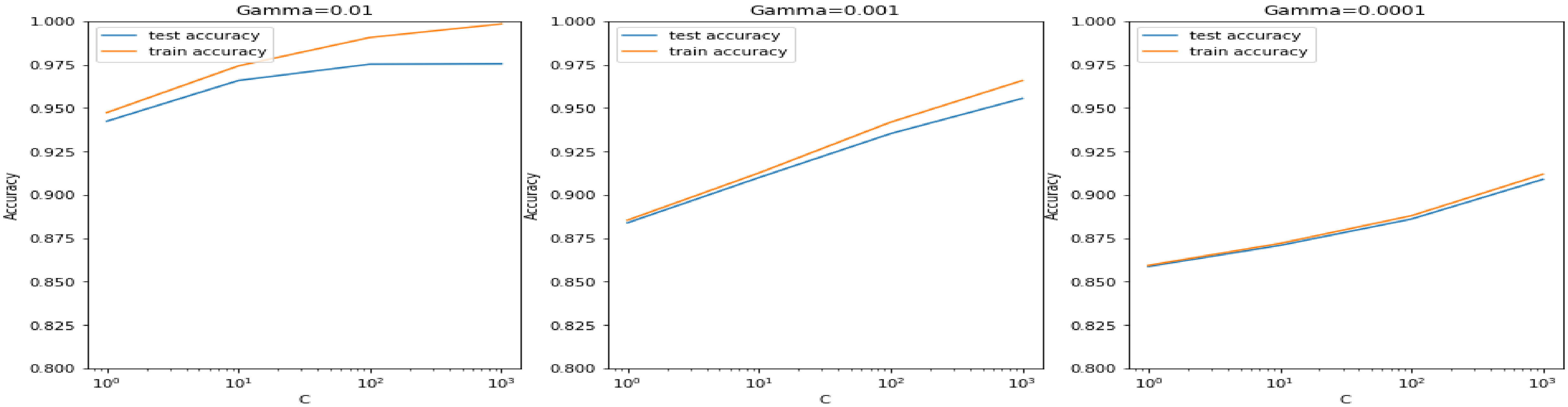


Model summary

Train set	Test set
Accuracy =0.86	Accuracy = 0.83
Sensitivity = 0.89	Sensitivity = 0.81
Specificity = 0.83	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



From the above plot, we can see that higher value of gamma leads to overfitting the model. With the lowest value of gamma (0.0001) we have train and test accuracy almost same.

Also, at C=100 we have a good accuracy and the train and test scores are comparable.
Though sklearn suggests the optimal scores mentioned above (gamma=0.01, C=1000), one could argue that it is better to choose a simpler, more non-linear model with gamma=0.0001. This is because the optimal values mentioned here are calculated based on the average test accuracy (but not considering subjective parameters such as model complexity).
We can achieve comparable average test accuracy (~90%) with gamma=0.0001 as well, though we'll have to increase the cost C for that. So to achieve high accuracy, there's a tradeoff between:
High gamma (i.e. high non-linearity) and average value of C
Low gamma (i.e. less non-linearity) and high value of C
We argue that the model will be simpler if it has as less non-linearity as possible, so we choose gamma=0.0001 and a high C=100.

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

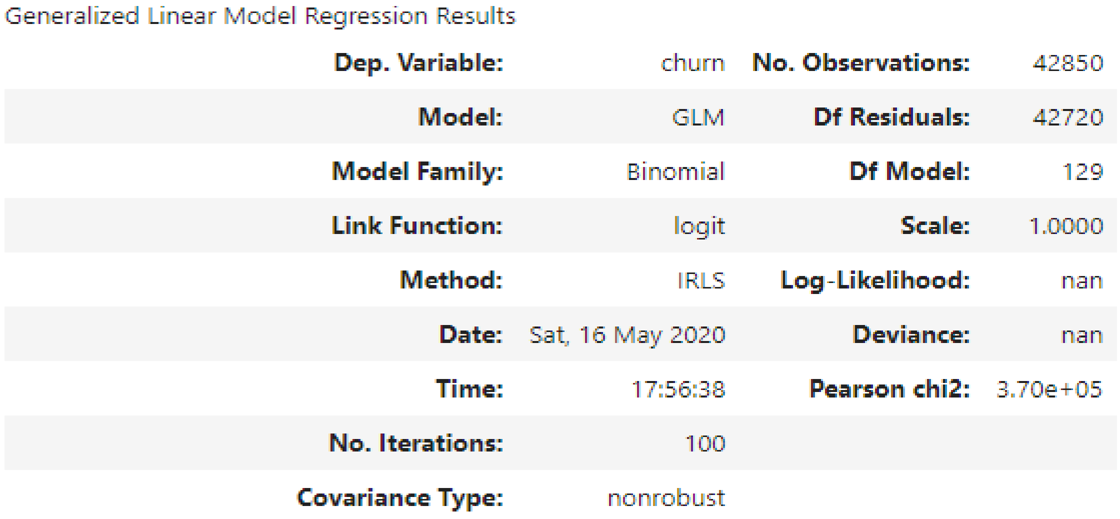
Specificity = 0.80

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%.

Logistic regression with No PCA

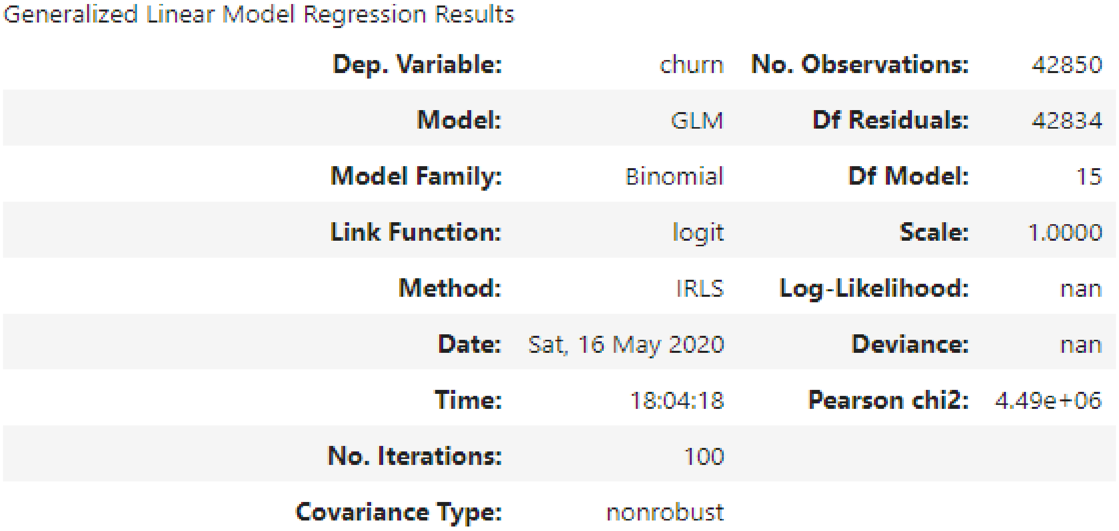


Generalized Linear Model Regression Results

Dep. Variable:	churn	**No. Observations:**	42850
Model:	GLM	**Df Residuals:**	42720
Model Family:	Binomial	**Df Model:**	129
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	nan
Date:	Sat, 16 May 2020	**Deviance:**	nan
Time:	17:56:38	**Pearson chi2:**	3.70e+05
No. Iterations:	100		
Covariance Type:	nonrobust		

Feature Selection Using RFE

Model-1 with RFE selected columns



Generalized Linear Model Regression Results

Dep. Variable:	churn	**No. Observations:**	42850
Model:	GLM	**Df Residuals:**	42834
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	nan
Date:	Sat, 16 May 2020	**Deviance:**	nan
Time:	18:04:18	**Pearson chi2:**	4.49e+06
No. Iterations:	100		
Covariance Type:	nonrobust		

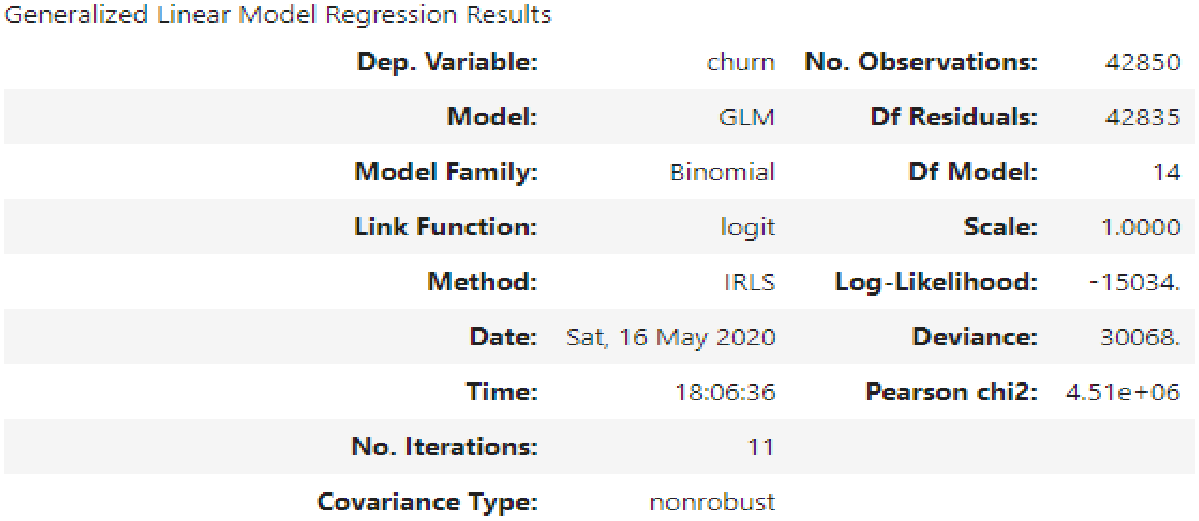
Model analysis

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).

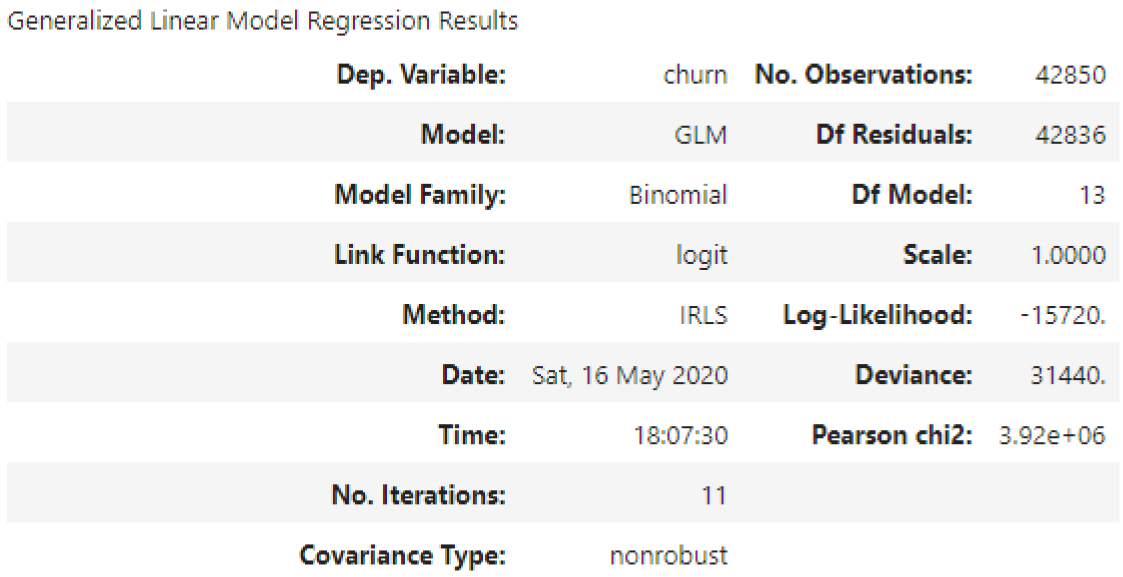
Model-2:- Building the model after removing og_others_8 variable.



Generalized Linear Model Regression Results

Dep. Variable:	churn	**No. Observations:**	42850
Model:	GLM	**Df Residuals:**	42835
Model Family:	Binomial	**Df Model:**	14
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-15034.
Date:	Sat, 16 May 2020	**Deviance:**	30068.
Time:	18:06:36	**Pearson chi2:**	4.51e+06
No. Iterations:	11		
Covariance Type:	nonrobust		

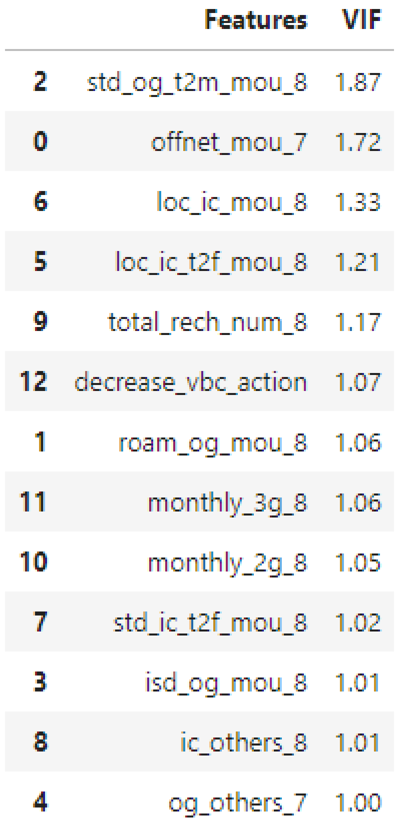
Model-3:- Model after removing offnet_mou_8 column.



Generalized Linear Model Regression Results

Dep. Variable:	churn	**No. Observations:**	42850
Model:	GLM	**Df Residuals:**	42836
Model Family:	Binomial	**Df Model:**	13
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-15720.
Date:	Sat, 16 May 2020	**Deviance:**	31440.
Time:	18:07:30	**Pearson chi2:**	3.92e+06
No. Iterations:	11		
Covariance Type:	nonrobust		

VIF Model-3

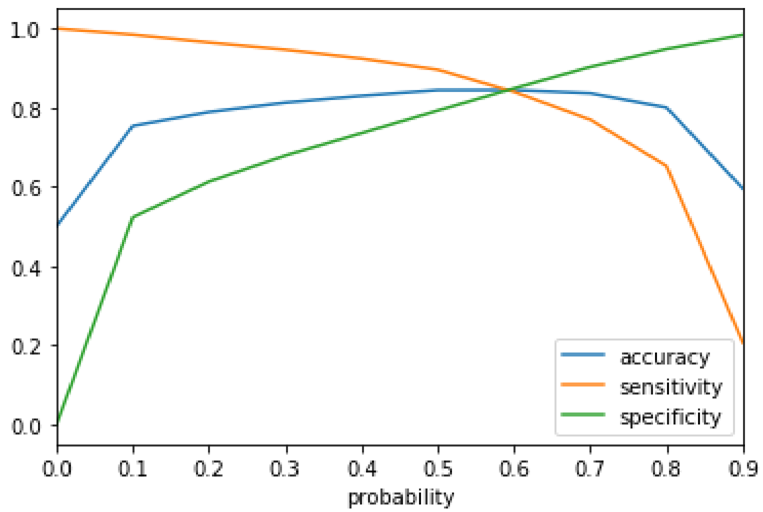


	Features	**VIF**
2	std_og_t2m_mou_8	1.87
0	offnet_mou_7	1.72
6	loc_ic_mou_8	1.33
5	loc_ic_t2f_mou_8	1.21
9	total_rech_num_8	1.17
12	decrease_vbc_action	1.07
1	roam_og_mou_8	1.06
11	monthly_3g_8	1.06
10	monthly_2g_8	1.05
7	std_ic_t2f_mou_8	1.02
3	isd_og_mou_8	1.01
8	ic_others_8	1.01
4	og_others_7	1.00

Now 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.

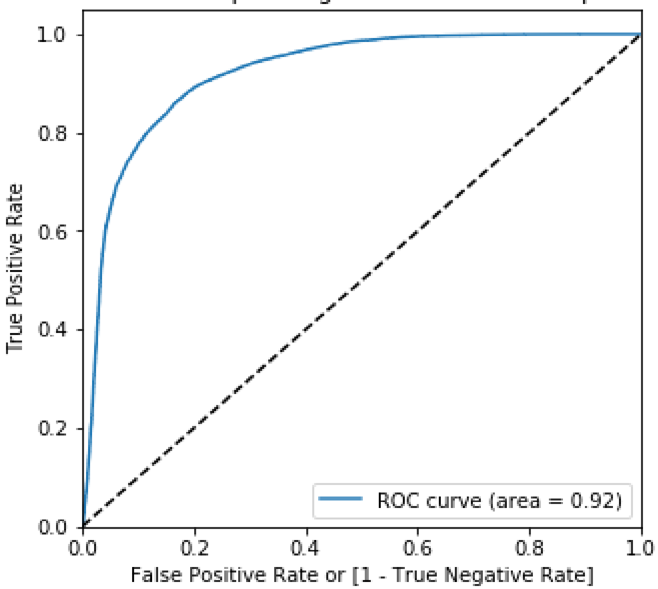
Hence, we can conclused that ***Model-3 log_no_pca_3 will be the final model***.

Model performance on the train set



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

Receiver operating characteristic example



Analysis of the above 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.

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

Model summary

Train set	Test set
Accuracy =0.84	Accuracy = 0.78
Sensitivity = 0.81	Sensitivity = 0.82
Specificity = 0.83	Specificity = 0.78

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

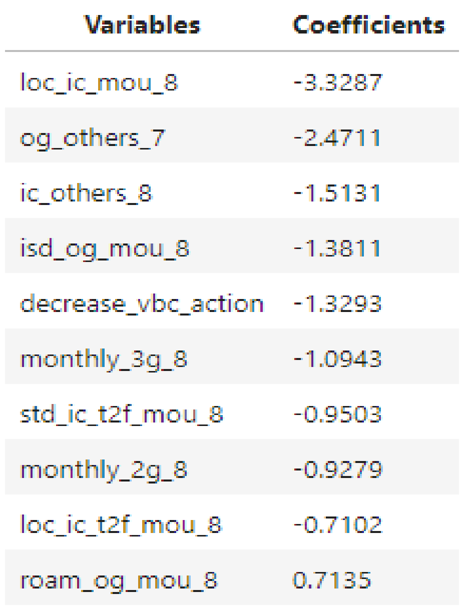
Final conclusion with no PCA

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 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 recommendation

Top predictors

Below are few top variables selected in the logistic regression model.



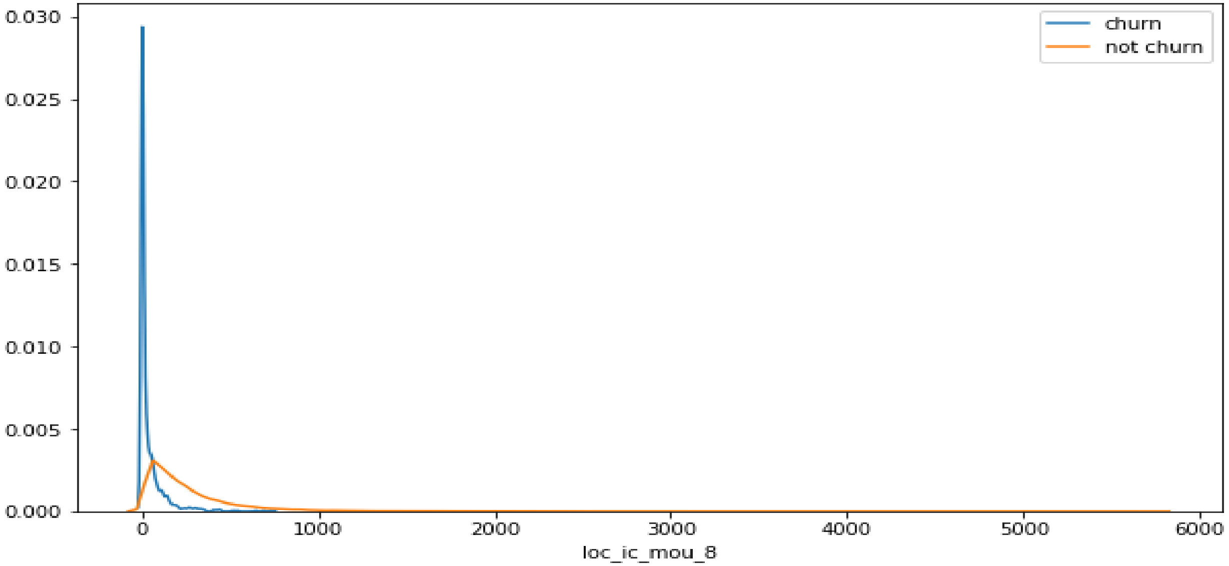
Variables	Coefficients
loc_ic_mou_8	-3.3287
og_others_7	-2.4711
ic_others_8	-1.5131
isd_og_mou_8	-1.3811
decrease_vbc_action	-1.3293
monthly_3g_8	-1.0943
std_ic_t2f_mou_8	-0.9503
monthly_2g_8	-0.9279
loc_ic_t2f_mou_8	-0.7102
roam_og_mou_8	0.7135

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.

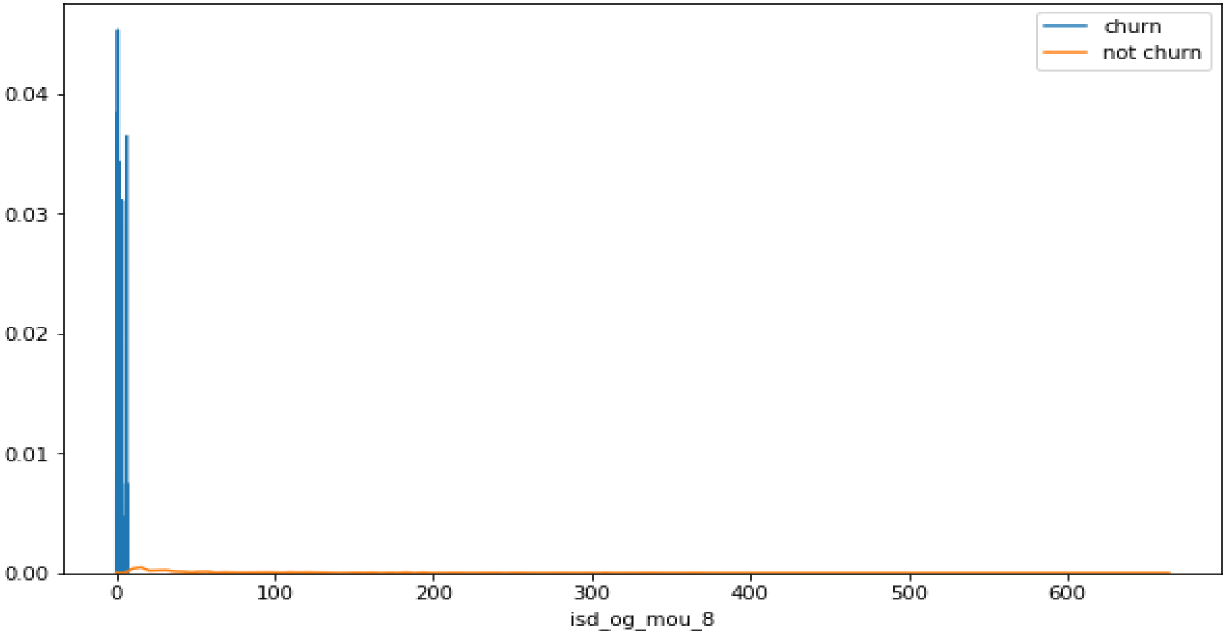
Recomendations

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. Cutomers, whose monthly 3G recharge in August is more, are likely to be churned.
5. 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.
6. Cutomers decreasing monthly 2g usage for August are most probable to churn.
7. Customers having decreasing incoming minutes of usage for operators T to fixed lines of T for August are more likely to churn.
8. 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.

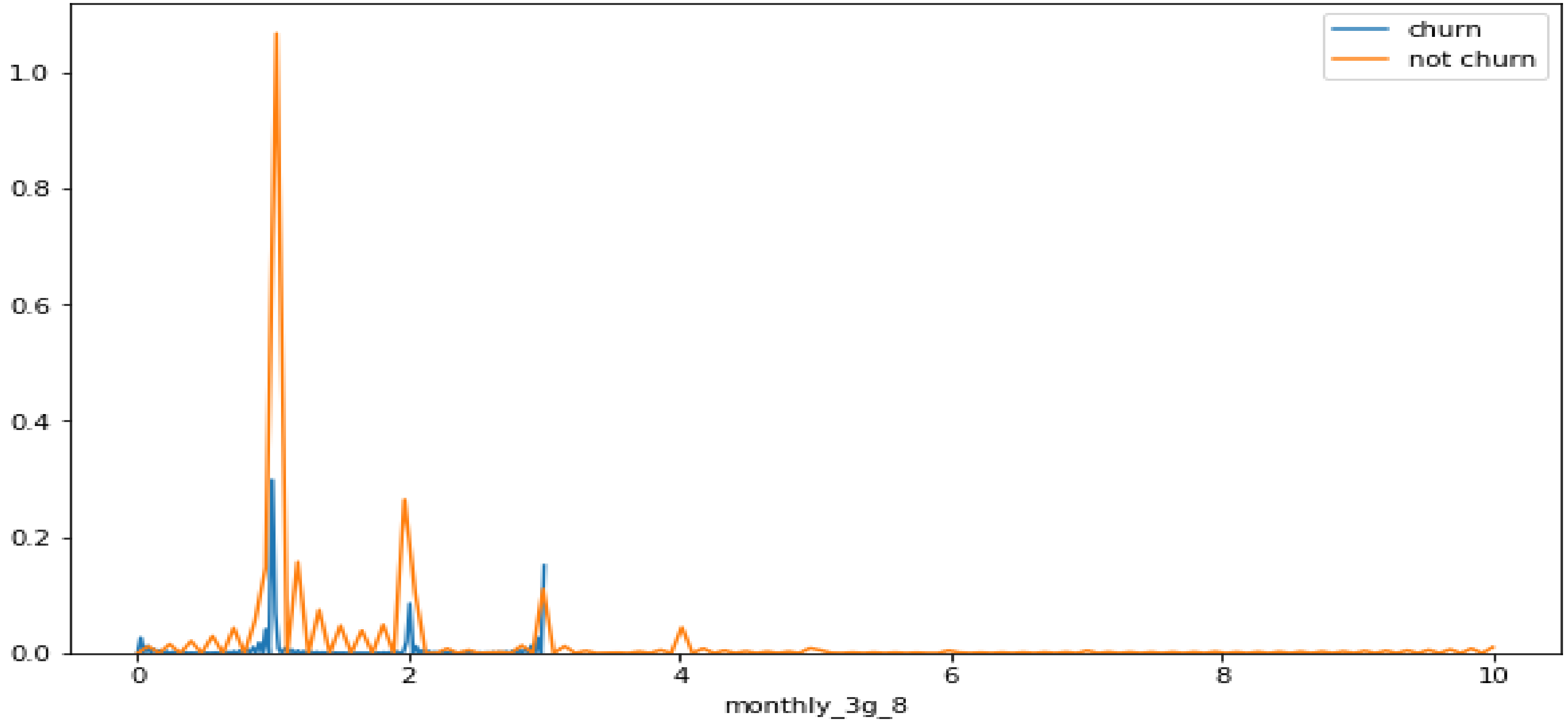
Plots of important predictors for churn and non churn customers



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 densed approximately to zero. On the onther hand for the non churn customers it is little more than the 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.