Telecom Churn Case Study– VIRAL SHAH

Problem Statement: To reduce customer churn, telecom companies need to predict which customers are at high risk of churn. In this activity, 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. Retaining high profitable customers is the main business goal here.

Business Objective: 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.

Steps Followed:

- ✓ Reading, Understanding and Visualizing the data
- ✓ Preparing the data for modelling
- ✓ Building the model
- ✓ Evaluating the model
- ✓ Analysing various models to find out the right model

Telecom Churn Case Study– VIRAL SHAH

- **Cleaning the dataset:** Finding missing values, Null values and identifying the no# of entries. If the entries are more, than better approach is to replace it with either Median values for continuous variables. And with mode for categorical variables
- **Columns Clean-ups**: Identifying columns which has more Null or blank values, than removing those columns. Also trying to understand if the columns are really necessary for the analysis.
- **Identifying outliers:** By plotting various plot, we tried to identify which are the outliers in the dataset. Addressing those outliers and generating the plots again. The idea is to create a date-set where we can measure the values.
- **Plotting columns:** By plotting various columns we analysed different trend in the data. We tried to pivot few columns to identify and understand the various trends.
- **Imbalance Data:** For the analysis, I also tried to understand the imbalance of data. Details are listed in the slides.

Telecom Churn Case Study– VIRAL SHAH

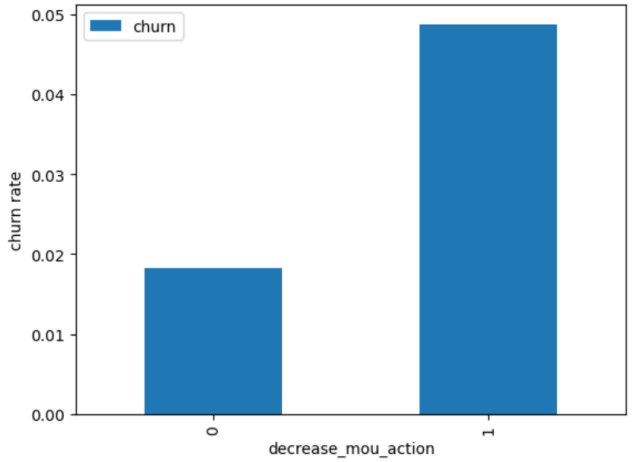
Started with Filtering high-value customers:
Creating column avg_rech_amt_6_7 by summing up total recharge amount of month 6 and 7. Then taking the average of the sum.

Tag Churners:
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. Removed attributes which are corresponding to churn phase.

Derived new columns:
- Total mou at good phase incoming and outgoing
- Avg rech number at action phase
- Avg rech_amt in action phase
- ARUP in action phase
- VBC in action phase

Telecom Churn Case Study– VIRAL SHAH

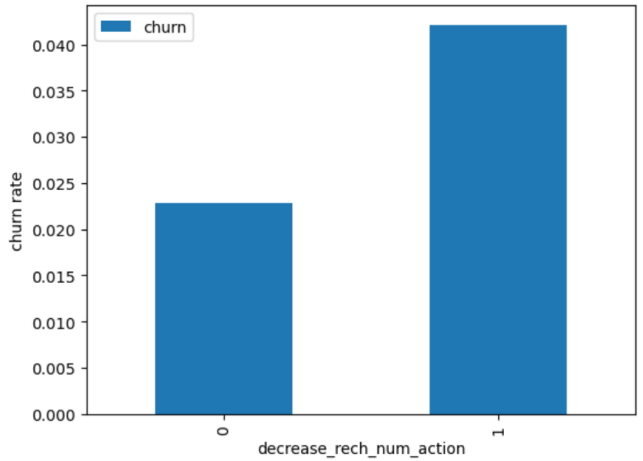
Univariate analysis



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.

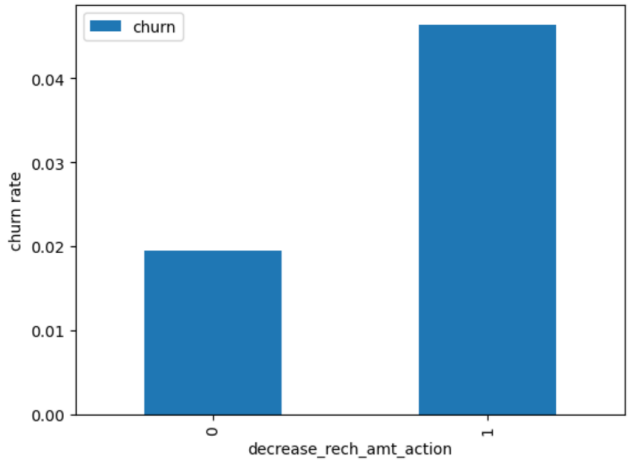
Churn rate on the basis whether the customer decreased her/his number of recharge in action month



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



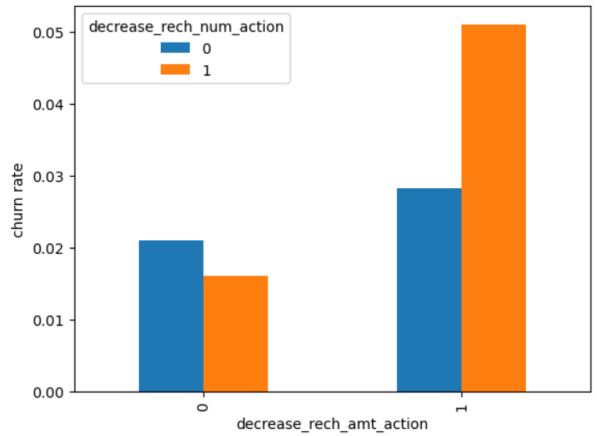
Analysis

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.

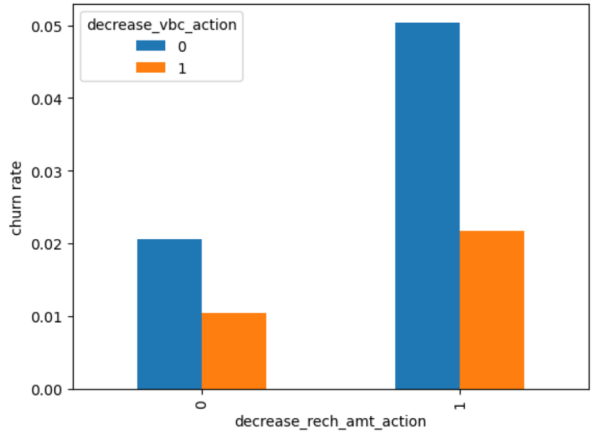
Churn rate on the basis whether the customer decreased her/his volume based cost in action month

Telecom Churn Case Study– VIRAL SHAH

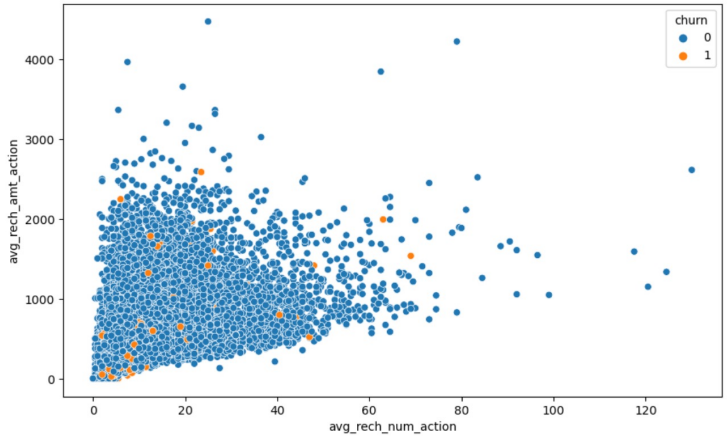
Bi-variate analysis



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



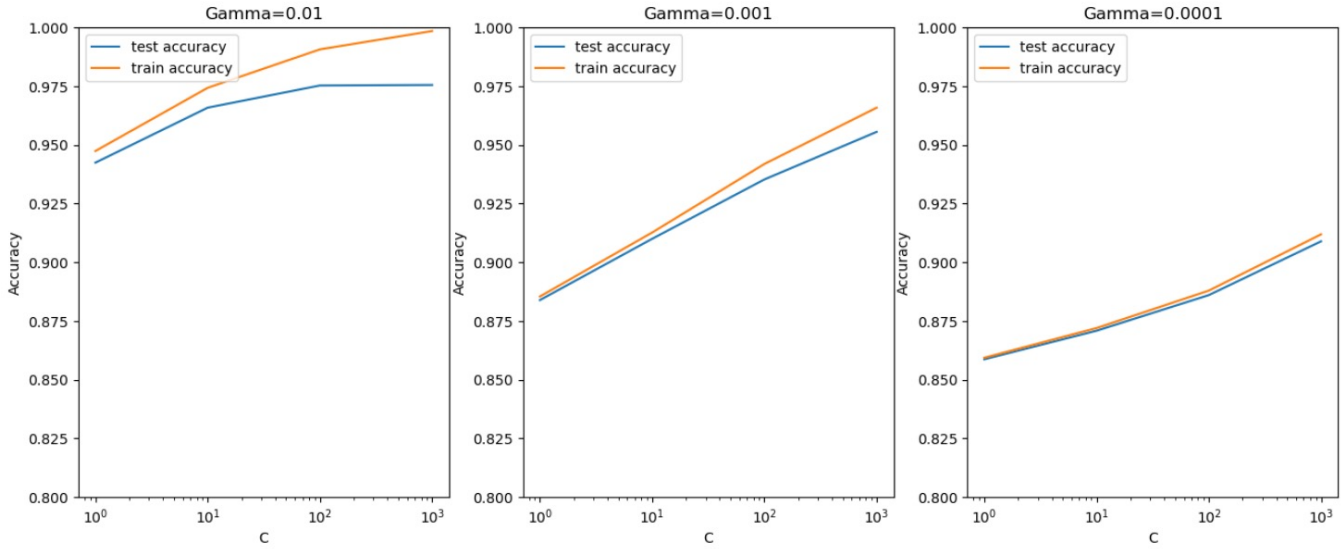
Analysis
Here, also we can see that the churn rate is more for the customers, whose recharge amount is decreased along with the volume based cost is increased in the action month.
Analysis of recharge amount and number of recharge in action month



Analysis
We can see from the above pattern that the recharge number and the recharge amount are mostly proportional. More the number of recharge, more the amount of the recharge.
Dropping few derived columns, which are not required in further analysis

Telecom Churn Case Study– VIRAL SHAH

Support Vector Machine(SVM) with PCA:



The best test score is **0.9754959911159373** corresponding to hyperparameters {'C': 1000, 'gamma': 0.01}
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.

Telecom Churn Case Study– VIRAL SHAH

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:	Thu, 02 Nov 2023	**Deviance:**	23572.
Time:	15:47:26	**Pearson chi2:**	3.71e+05
No. Iterations:	100	**Pseudo R-squ. (CS):**	nan
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-57.0082	4420.570	-0.013	0.990	-8721.166	8607.150
loc_og_t2o_mou	3.071e-06	0.000	0.013	0.990	-0.000	0.000
std_og_t2o_mou	1.19e-06	9.96e-05	0.012	0.990	-0.000	0.000
loc_ic_t2o_mou	-1.749e-07	2.38e-07	-0.733	0.463	-6.42e-07	2.92e-07
arpu_6	-0.0338	0.081	-0.418	0.676	-0.192	0.125
arpu_7	0.0855	0.086	0.995	0.320	-0.083	0.254
arpu_8	0.0909	0.110	0.828	0.408	-0.124	0.306
onnet_mou_6	15.5141	3.579	4.335	0.000	8.500	22.528
onnet_mou_7	-4.3250	1.811	-2.389	0.017	-7.874	-0.776
onnet_mou_8	2.3520	1.828	1.287	0.198	-1.231	5.935
offnet_mou_6	15.0883	3.366	4.482	0.000	8.490	21.686

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.

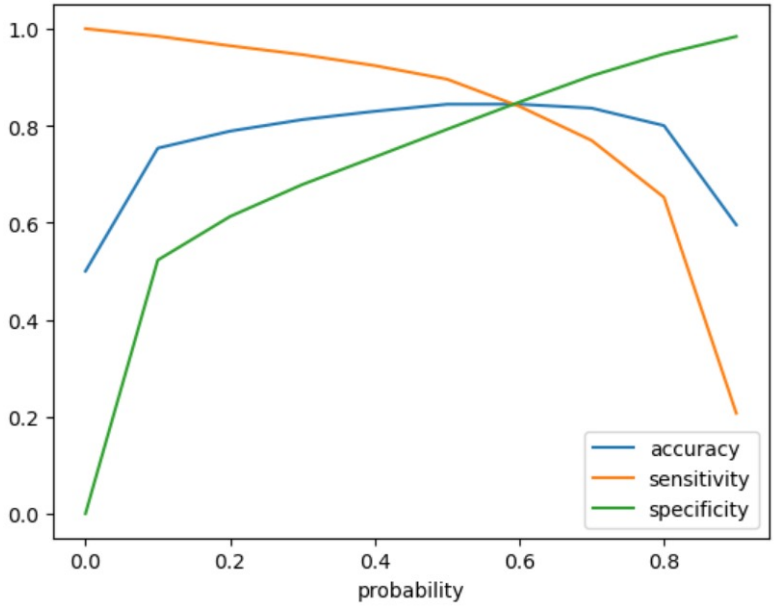
Coarse tuning (Auto+Manual)

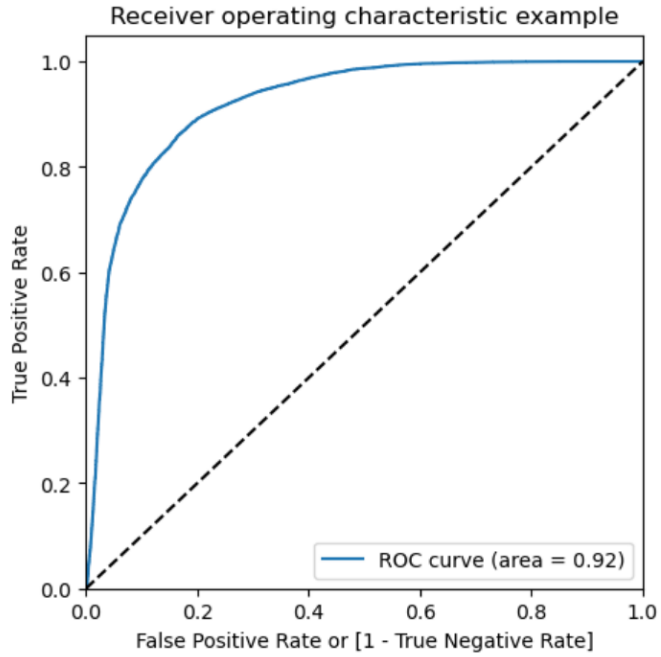
We'll first eliminate a few features using Recursive Feature Elimination (RFE), and once we have reached a small set of variables to work with, we can then use manual feature elimination (i.e. manually eliminating features based on observing the p-values and VIFs).

Telecom Churn Case Study– VIRAL SHAH

Building model and checking VIF

	probability	accuracy	sensitivity	specificity
0.0	0.0	0.500000	1.000000	0.000000
0.1	0.1	0.753629	0.984411	0.522847
0.2	0.2	0.788751	0.964714	0.612789
0.3	0.3	0.812509	0.946371	0.678646
0.4	0.4	0.829638	0.923874	0.735403
0.5	0.5	0.844131	0.895823	0.792439
0.6	0.6	0.844271	0.839860	0.848681
0.7	0.7	0.836173	0.769522	0.902824
0.8	0.8	0.800163	0.652275	0.948051
0.9	0.9	0.595426	0.207001	0.983851





Analysis of the above curve

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 cut-off, we are taking *0.5* for achieving higher sensitivity, which is our main goal.

Telecom Churn Case Study– VIRAL SHAH

Conclusion

Accuracy:- 0.7848763761053962
Sensitivity:- 0.8238341968911918
Specificity:- 0.7834704562453254

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.

Telecom Churn Case Study– VIRAL SHAH

Conclusion

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

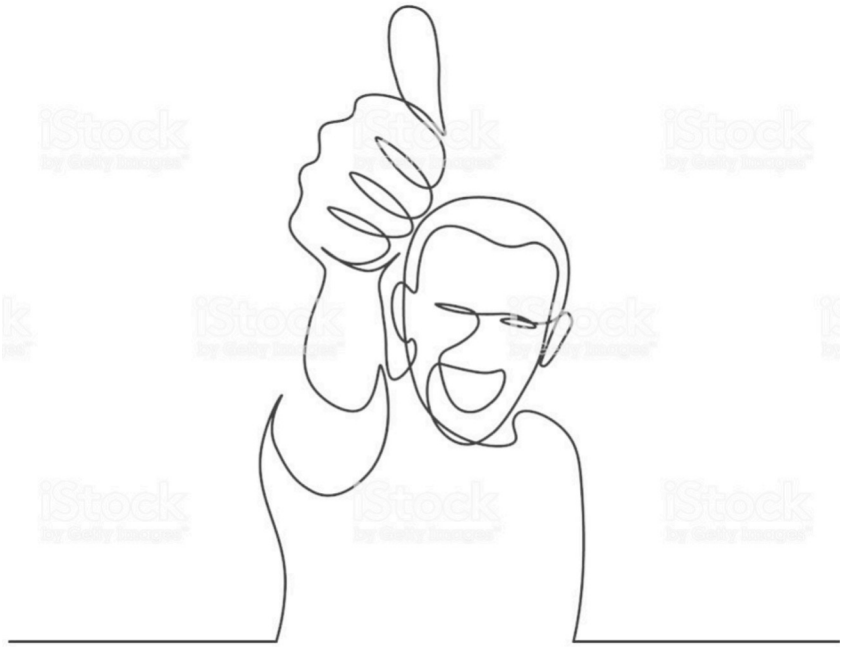
We can see most of the top variables have negative coefficients. That means, the variables are inversely correlated with the churn probability.

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.

Telecom Churn Case Study– VIRAL SHAH

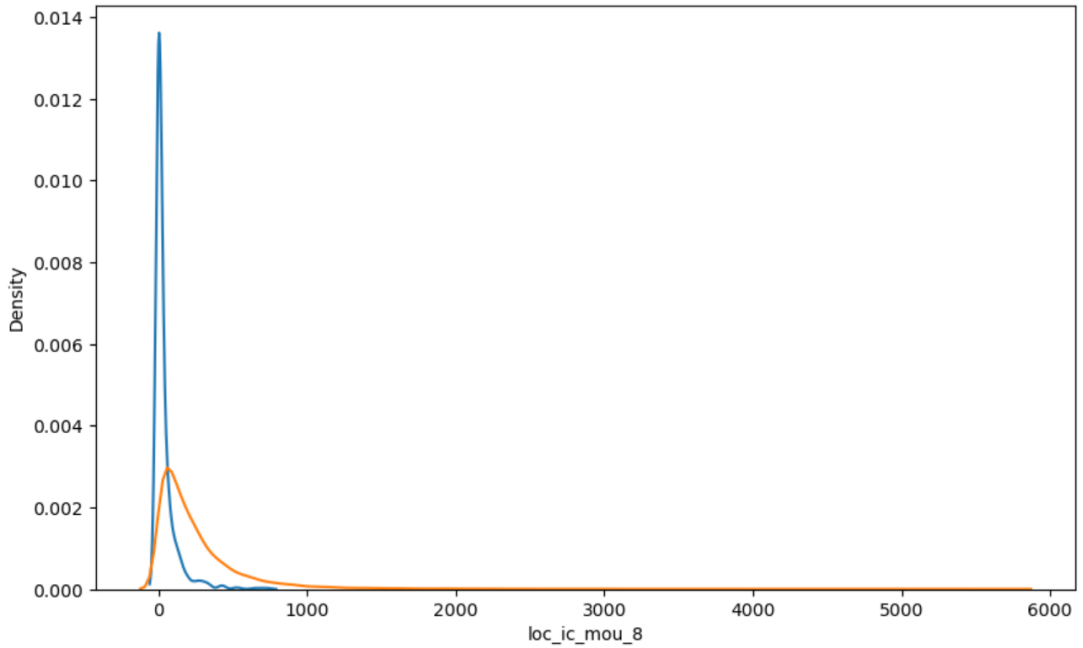
Recommendations

- 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 (0.7135). That means for the customers, whose roaming outgoing minutes of usage is increasing are more likely to churn.

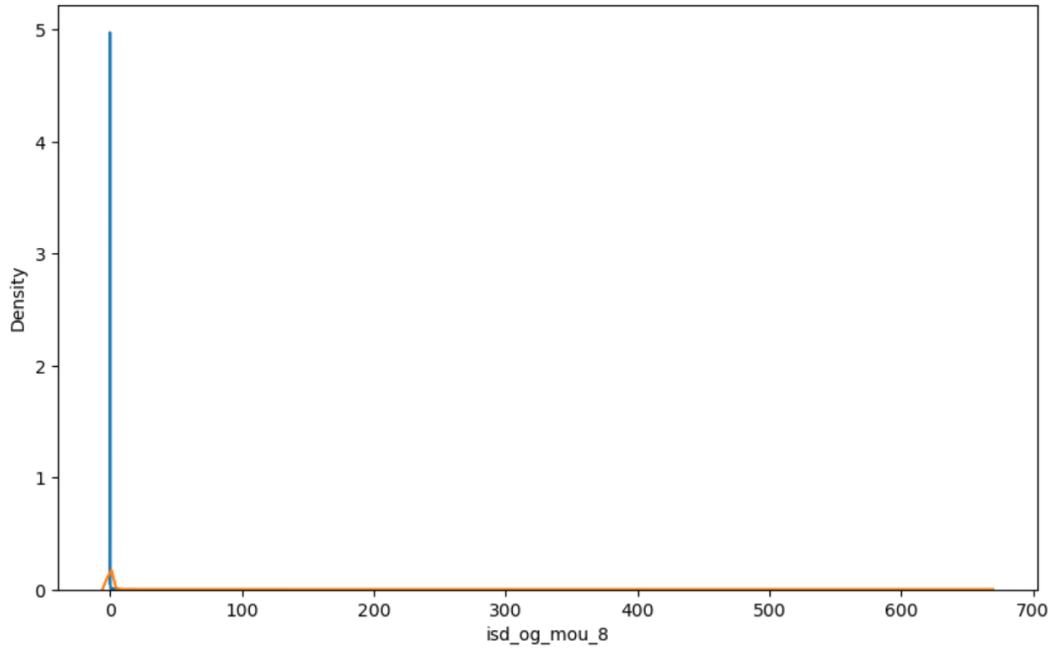


Telecom Churn Case Study– VIRAL SHAH

Plots of important predictors for churn and non churn customers



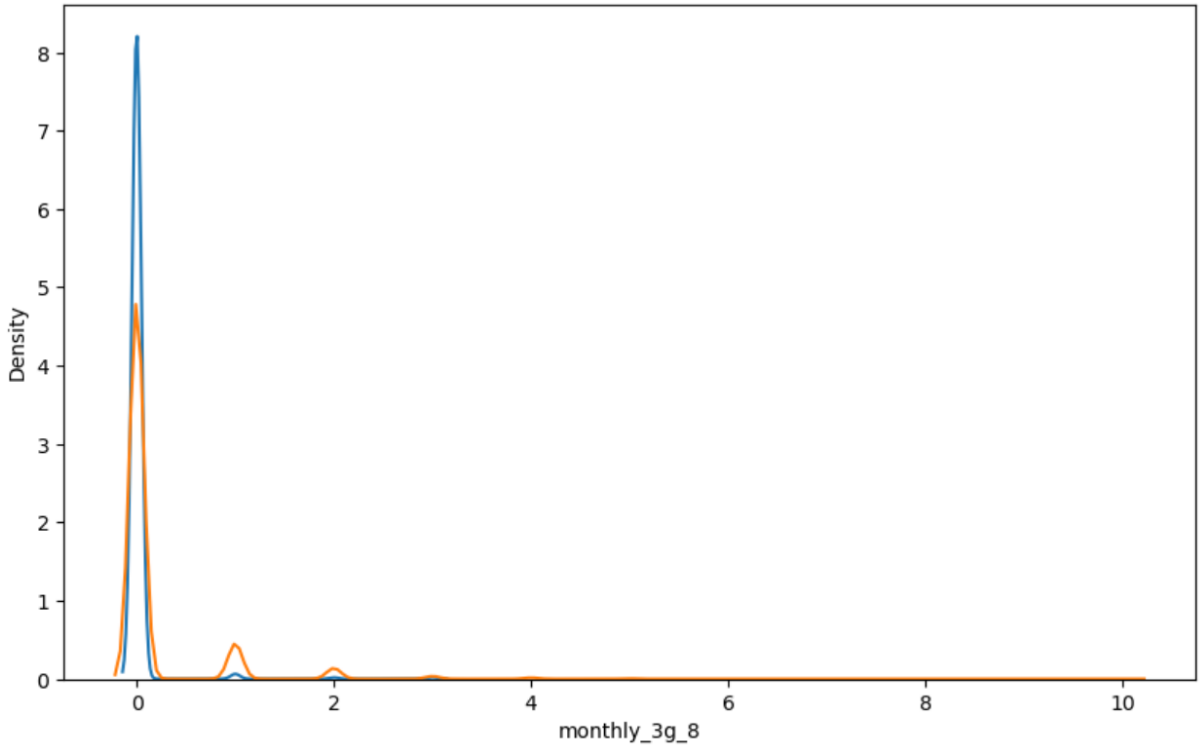
We can see that for the churn customers the minutes of usage for the month of August is mostly populated on the lower side than the non churn customers.



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

Telecom Churn Case Study– VIRAL SHAH

Plots of important predictors for churn and non 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 spreader across various numbers.

Similarly we can plot each variables, which have higher coefficients, churn distribution.

Telecom Churn Case Study– VIRAL SHAH

Thank You