Telecom Churn Case Study

Submitted by: Khushi Vora
Neha Vinod Nair
Gyanender Sindhu

Problem Statement

- Telecom industry characterized by high competition and annual churn rates of 15-25%.
- Acquiring new customers costs 5-10 times more than retaining existing ones, emphasizing the importance of customer retention.
- Retaining high-profit customers is a top priority for many incumbent operators.
- Key project goal: Analyze customer-level data of a leading telecom firm to build predictive models for identifying high-risk churn customers.
- Objective: Identify main indicators of churn to inform retention strategies.
- Approach:
- Analyze customer-level data.
- Build predictive models to identify high-risk churn customers.
- Importance of churn prediction:
- Allows proactive measures to retain customers.
- Helps optimize resource allocation for retention efforts.
- Outcome: Development of strategies focused on retaining high-value customers and reducing churn rates.

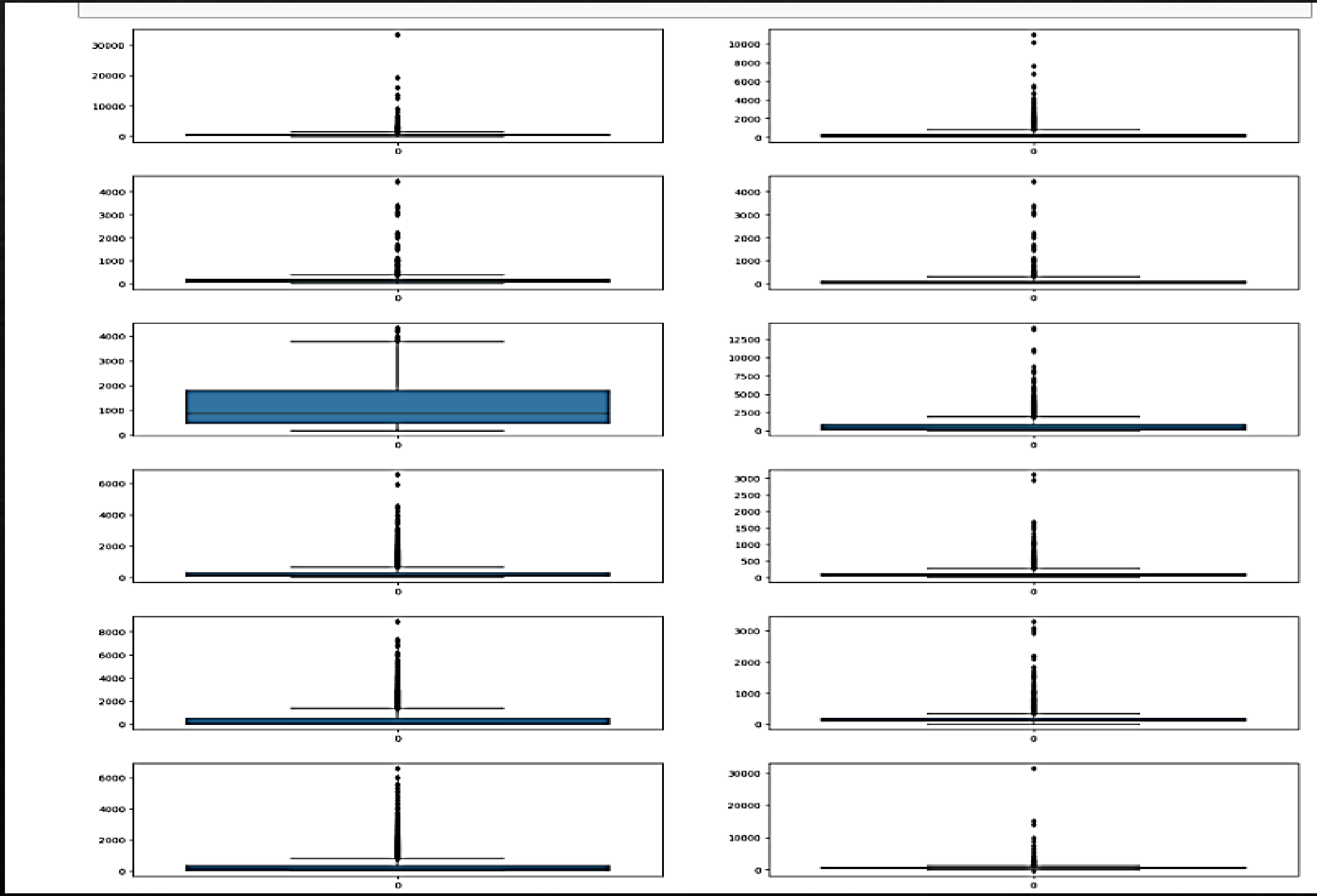
Objectives

- To predict Customer Churn rate
- Finding the main variables/factors influencing Customer Churn
- ML algorithms to build prediction models, evaluate the accuracy and performance of the models- Decision Tree, random forest, Logistic Regression.
- Finding out the best model for our business case and providing executive suggestions.

Steps

- Business Understanding
- Data Reading and Understanding
- EDA
- Data Preparation
- Building Models
- Validating the performace of Models
- Conclusion

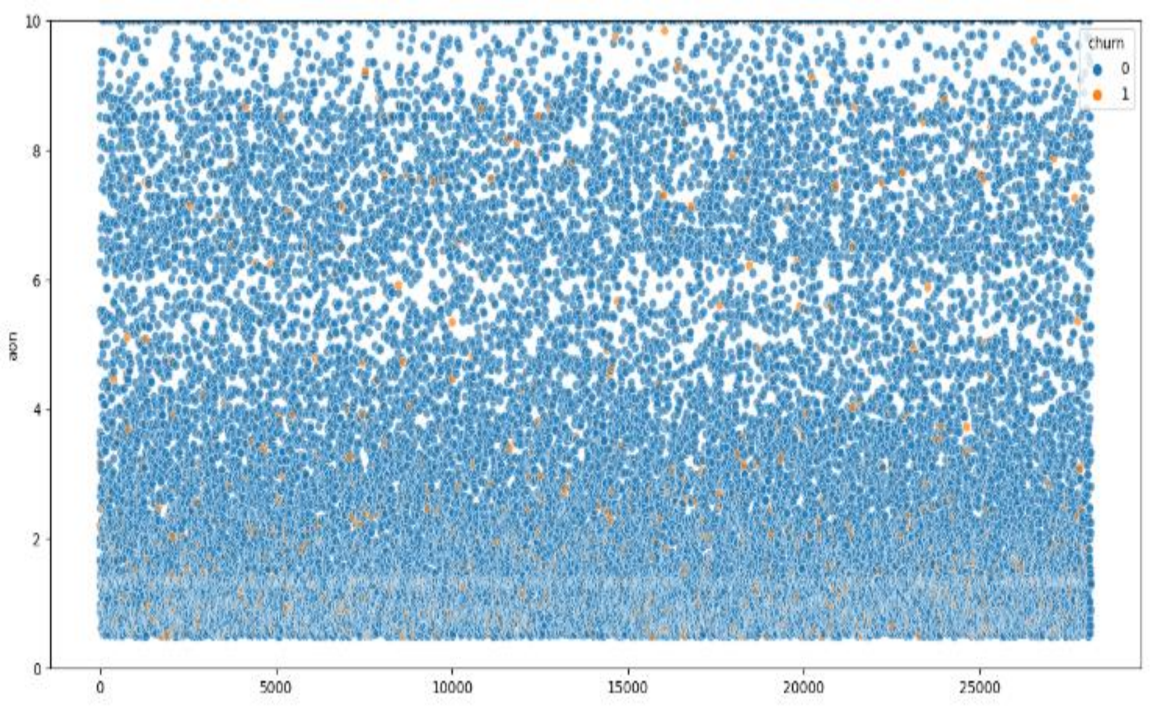
EDA



From the plots we can define following upper limits to the suspected variables

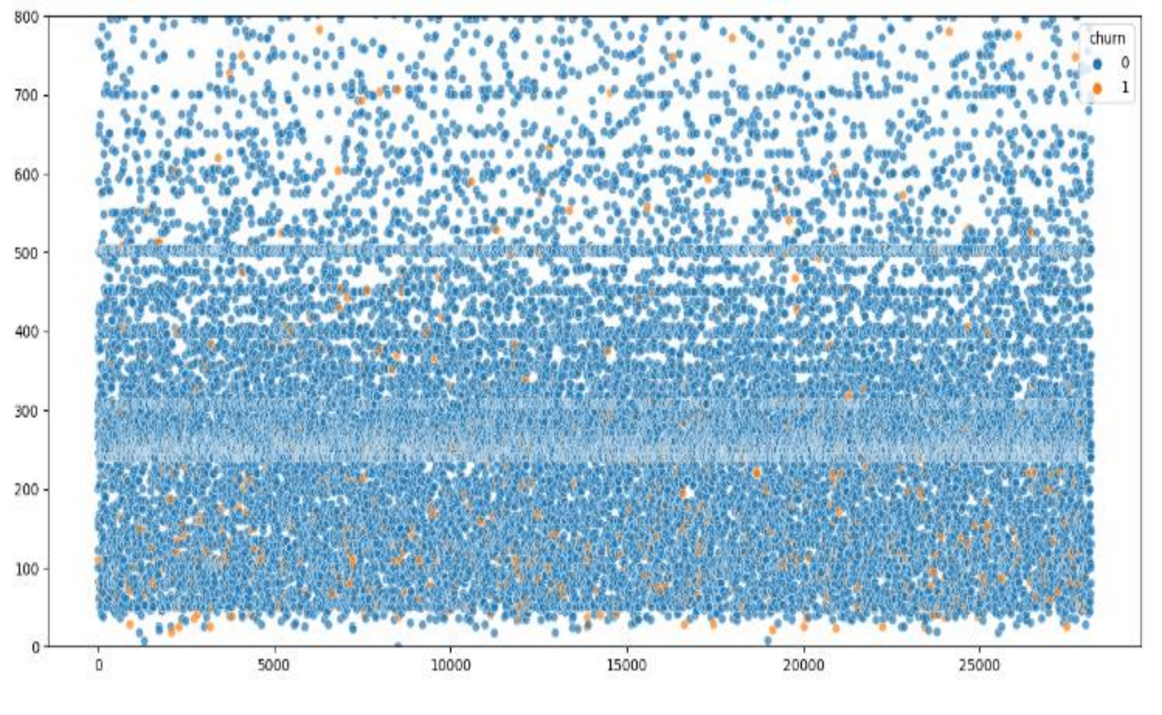
Feature	Value
arpu_8	7000
loc_og_mou_8	4000
max_rech_amt_8	1000
last_day_rch_amt_8	1000
aon	3000
total_mou_8	4000
gd_ph_loc_ic_mou	3000
gd_ph_last_day_rch_amt	1000
gd_ph_std_og_mou	4000
gd_ph_max_rech_amt	1500
gd_ph_loc_og_mou	3000
gd_ph_arpu	7000

Churn based on tenure



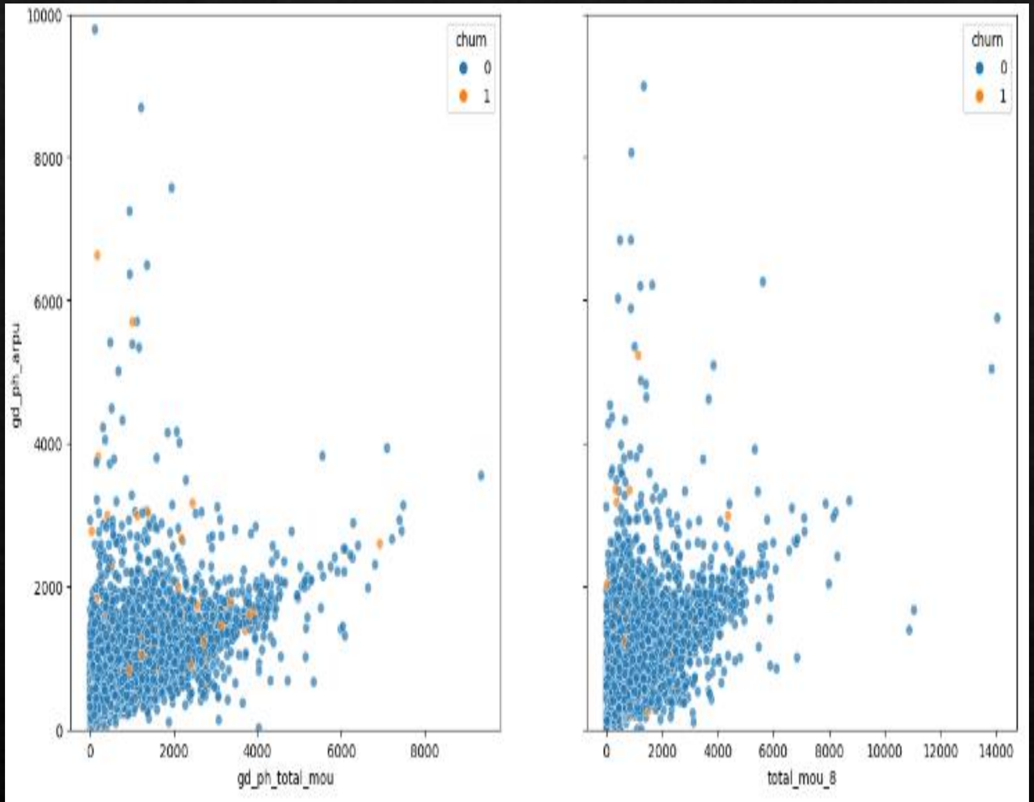
Though we cannot see a clear pattern here, but we can notice that the majority of churners had a tenure of less than 4 years

Effect of max recharge amount on churn



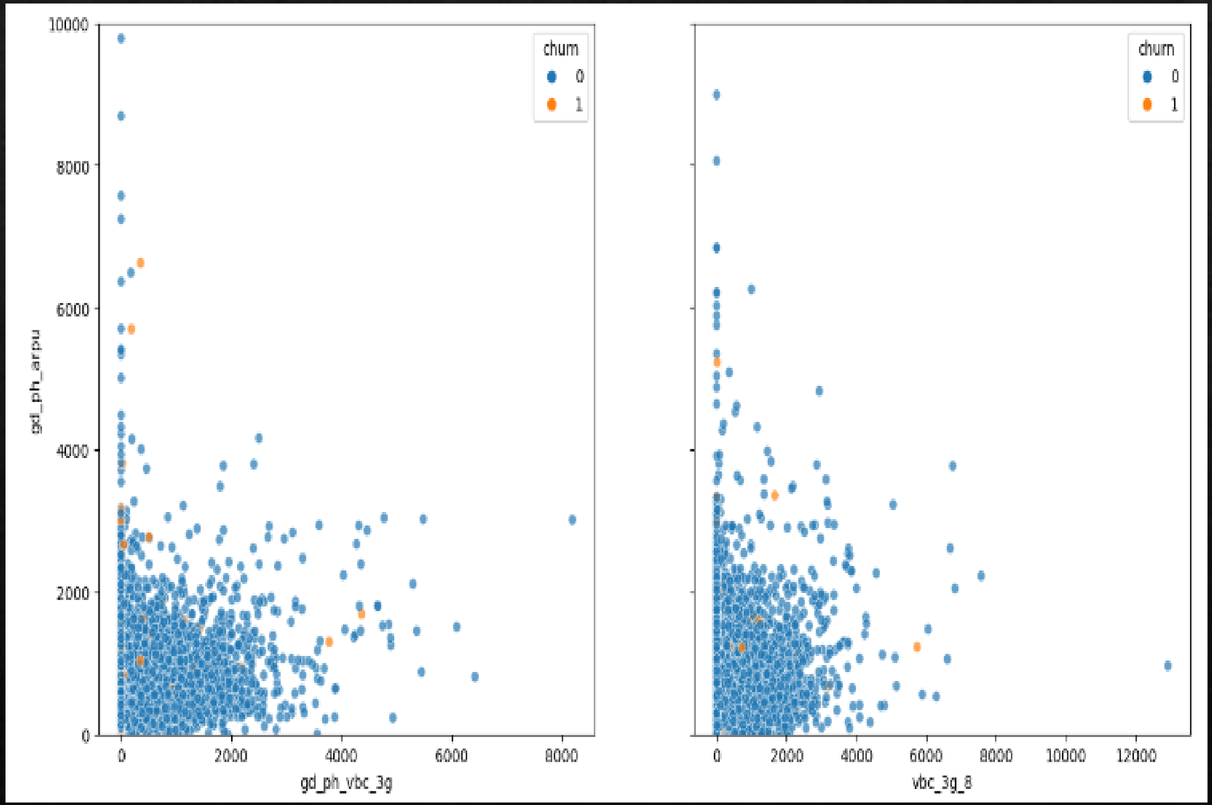
We can see that users who had the max recharge amount less the 200 churned more

VBC effects the revenue



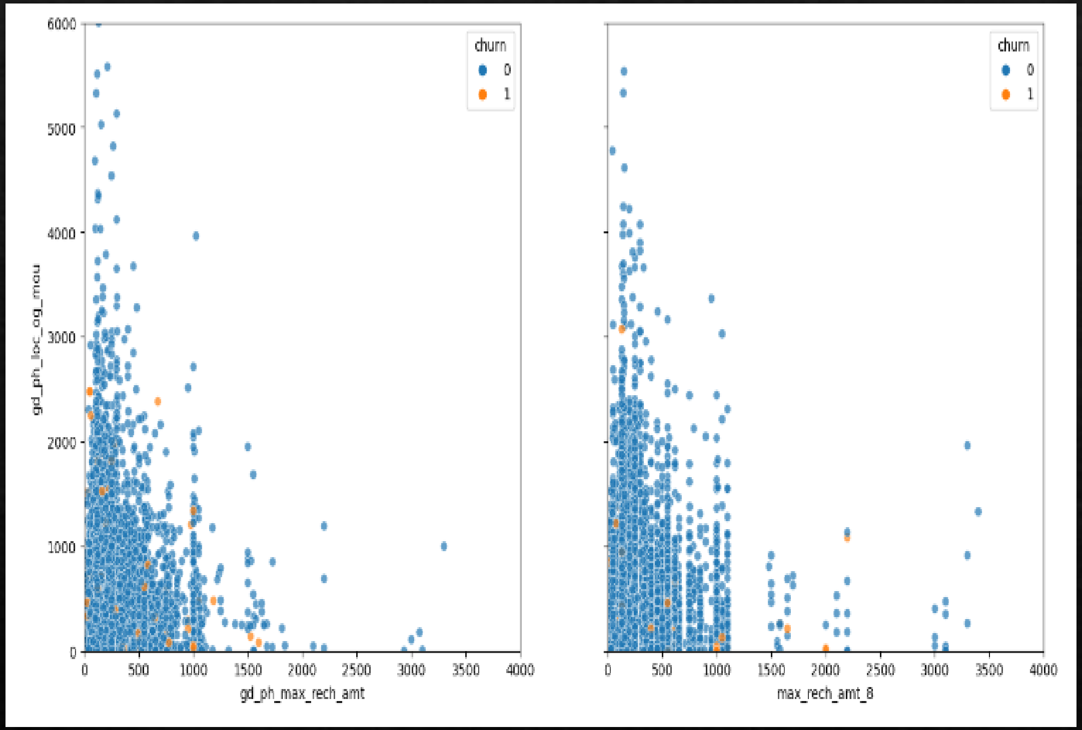
As we can observe that MOU is dropping significantly for churners in action phase which hitting the revenue generation. But then also revenue is higher in that part which indicates that the users are taking other services which increasing the revenue generation.

The total_mou effects the revenue



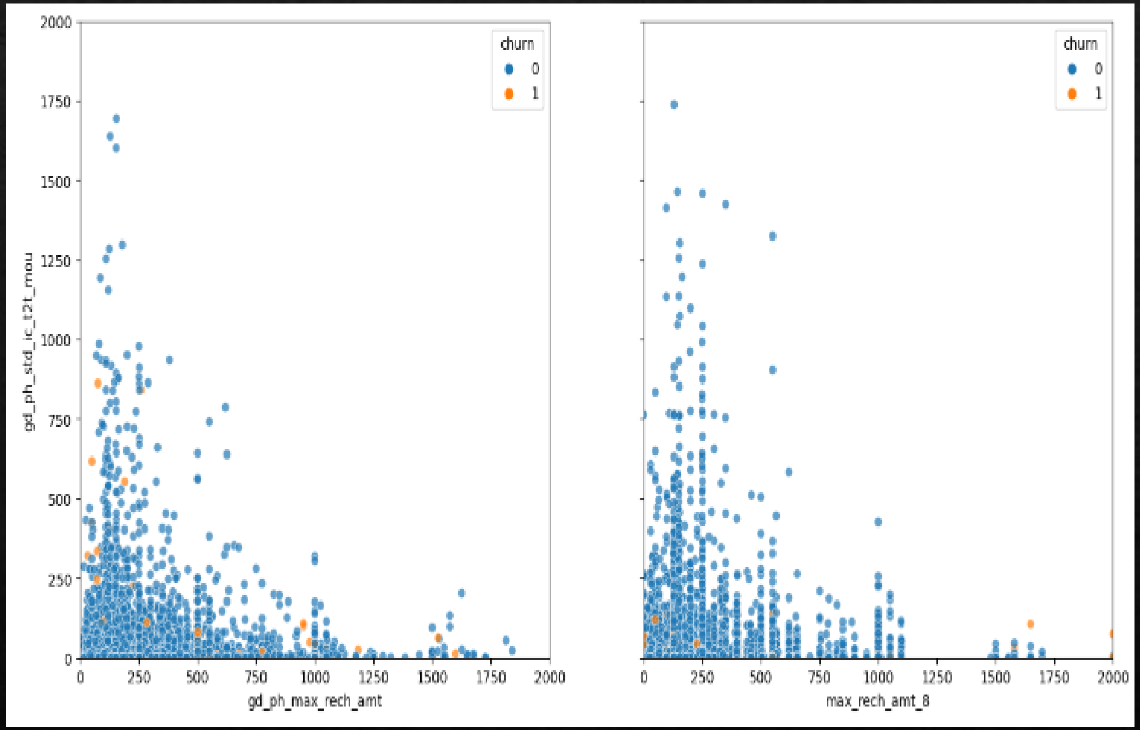
As we can see users using less amount of VBC generating high revenue churned and also revenue is higher from less consumption part.

Recharge amount vs local outgoing calls



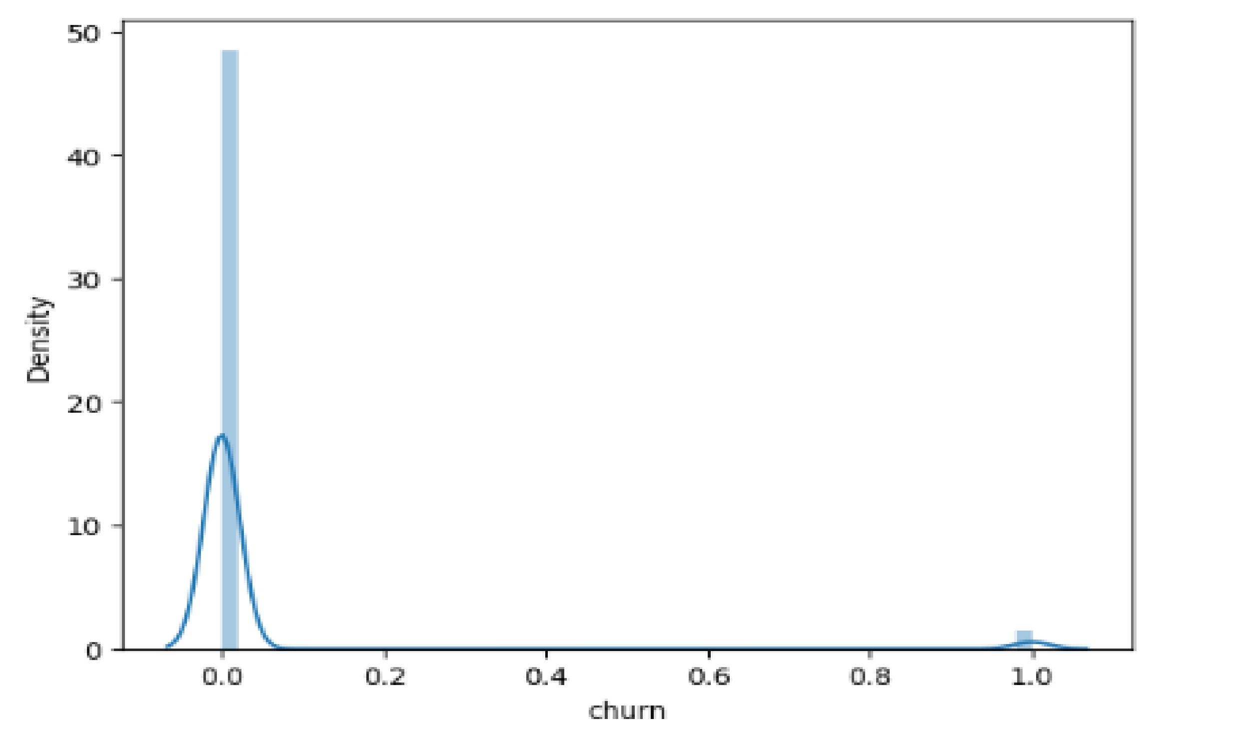
Users who were recharging with high amounts were using the service for local uses less as compared to user who did lesser amounts of recharge and people whose max recharge amount as well as local out going were very less even in the good phase churned more.

Service provider vs the recharge amount



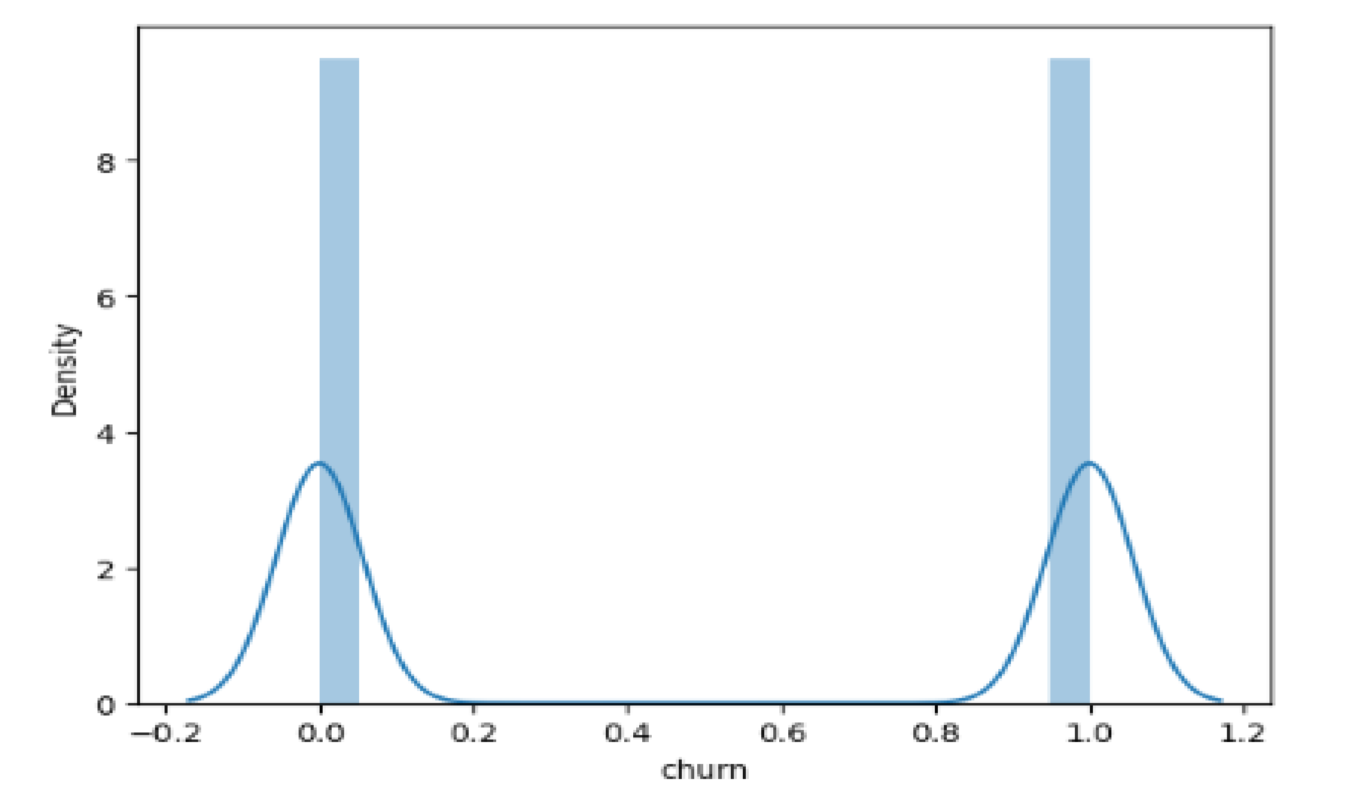
Users who have max recharge amount on the higher end and still have low incoming call mou during the good phase, churned out more

Distribution of target variable



As we can see that it is not skewed but highly imbalanced. The number of non churners are very high. so we will handle this using SMOTE.

Handling class imbalance using SMOTE



We can see now target is not skewed and class is balance.

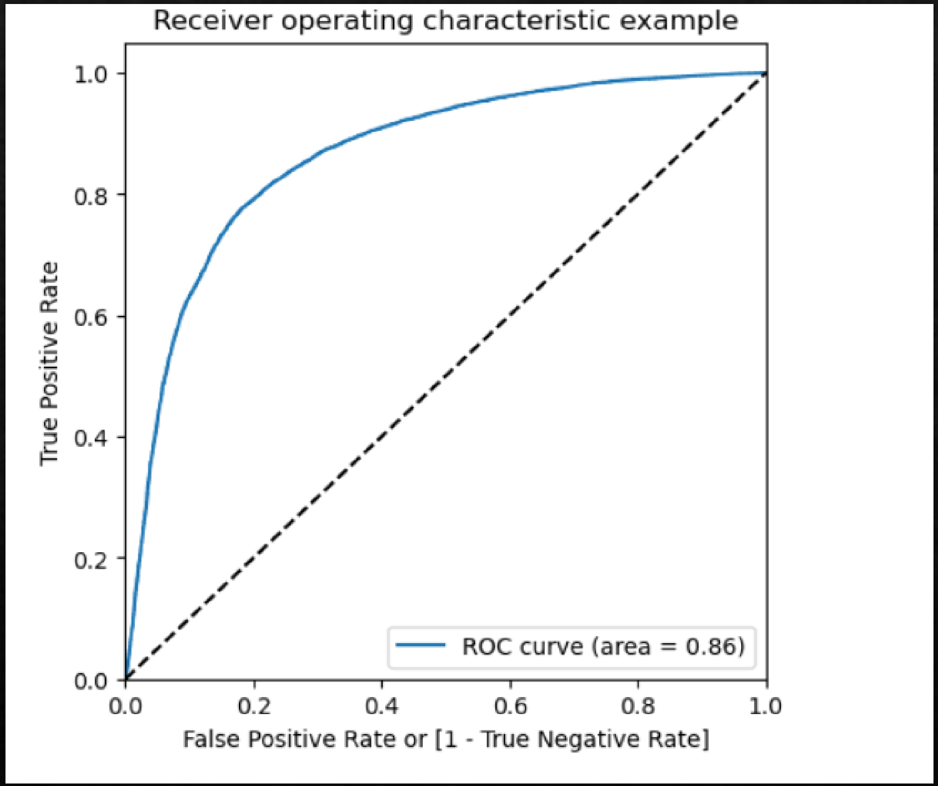
Model Building

1. Logistic Regression using RFE

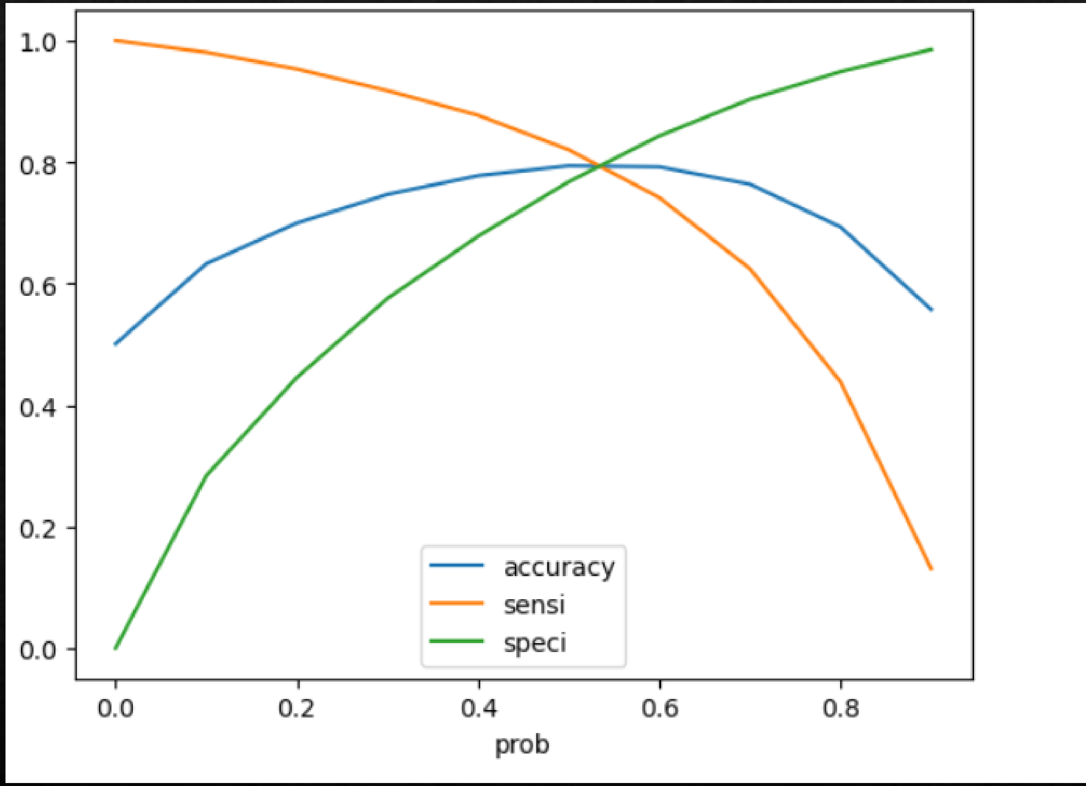
Dep. Variable:	churn	No. Observations:	38213
Model:	GLM	Df Residuals:	38157
Model Family:	Binomial	Df Model:	55
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-17495.
Date:	Tue, 07 May 2024	Deviance:	34989.
Time:	18:33:26	Pearson chi2:	1.39e+05
No. Iterations:	7	Pseudo R-squ. (CS):	0.3754
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3573	0.021	-63.458	0.000	-1.399	-1.315
arpu_8	0.3533	0.033	10.825	0.000	0.289	0.417
roam_ic_mou_8	-0.3624	0.026	-14.202	0.000	-0.412	-0.312
loc_og_mou_8	-0.2828	0.047	-6.008	0.000	-0.375	-0.191
loc_ic_mou_8	-1.7448	0.058	-30.105	0.000	-1.858	-1.631
std_ic_t2t_mou_8	-0.3962	0.042	-9.417	0.000	-0.479	-0.314
spl_ic_mou_8	-0.2286	0.021	-10.804	0.000	-0.270	-0.187
total_rech_num_8	-0.5703	0.032	-17.630	0.000	-0.634	-0.507
max_rech_amt_8	0.2382	0.022	10.779	0.000	0.195	0.282
last_day_rch_amt_8	-0.5497	0.021	-26.072	0.000	-0.591	-0.508
vol_2g_mb_8	-0.2671	0.030	-8.989	0.000	-0.325	-0.209
monthly_2g_8	-0.6972	0.025	-27.787	0.000	-0.746	-0.648
sachet_2g_8	-0.4703	0.023	-20.526	0.000	-0.515	-0.425
monthly_3g_8	-0.9591	0.036	-26.835	0.000	-1.029	-0.889
sachet_3g_8	-0.4200	0.035	-11.884	0.000	-0.489	-0.351
aon	-0.3985	0.016	-24.794	0.000	-0.430	-0.367
total_mou_8	-0.8328	0.037	-22.587	0.000	-0.905	-0.761
gd_ph_total_mou	-0.8290	0.203	-4.078	0.000	-1.227	-0.431
gd_ph_std_og_mou	1.0200	0.179	5.697	0.000	0.669	1.371
gd_ph_roam_og_mou	0.3068	0.033	9.295	0.000	0.242	0.371
gd_ph_monthly_3g	0.2872	0.023	12.505	0.000	0.242	0.332

ROC Plotting



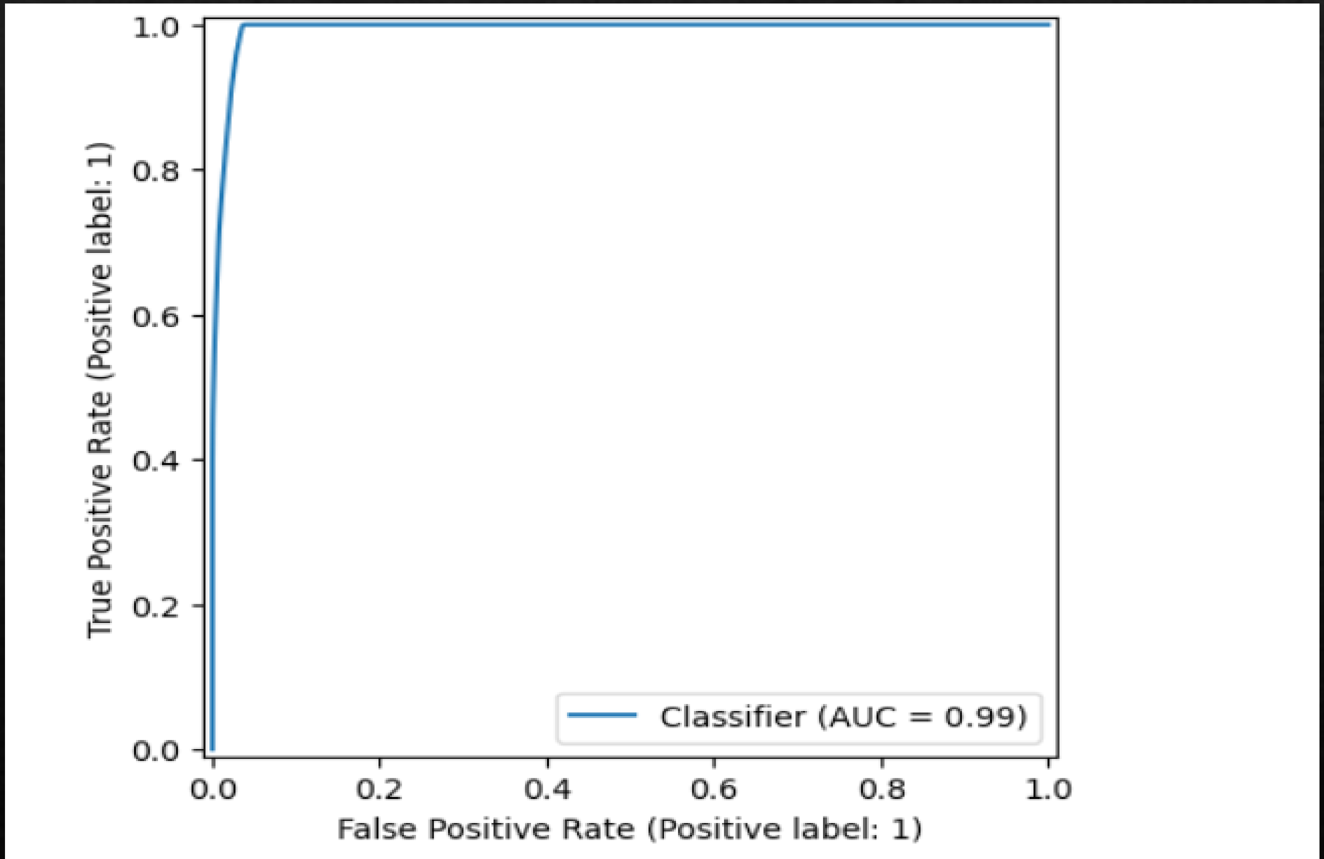
Optimal Cutoff



As we can see optimal cutoff is 0.5 so we will keep it

Decision Tree

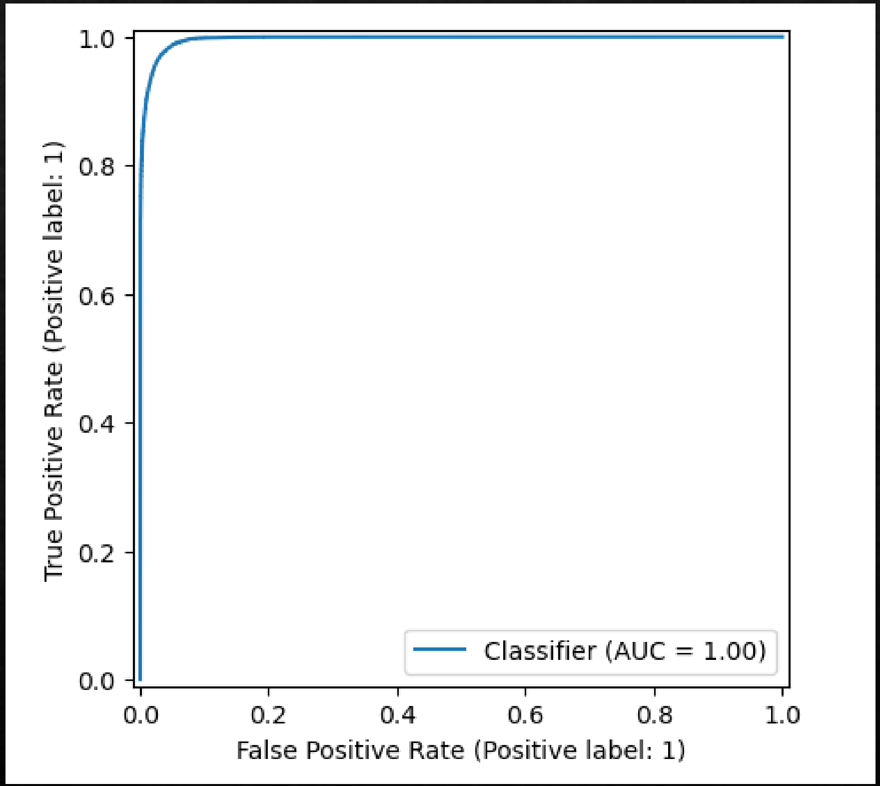
ROC curve using hyperparameter tunning



With Decision Tree, we are getting 89% accuracy.

Random Forest

ROC curve using hyperparameter tunning



With Random Forest, we are getting 94% accuracy

- **Business Insights**

1. In our effort to retain customers, prioritizing recall is crucial. Identifying customers at risk of churn (true positives) is paramount, as it allows us to intervene effectively and prevent customer loss. This approach minimizes the cost associated with losing a customer and acquiring new ones.
2. Upon evaluating the trained models, we've observed that the tuned Random Forest model outperforms others, boasting the highest accuracy and recall rates at 94%. Therefore, selecting the Random Forest model aligns with our objective of maximizing customer retention effectiveness.

Final Model

Report on train data

	precision	recall	f1-score	support
0	0.98	0.96	0.97	19080
1	0.96	0.98	0.97	19133
accuracy			0.97	38213
macro avg	0.97	0.97	0.97	38213
weighted avg	0.97	0.97	0.97	38213

Report on test data

	precision	recall	f1-score	support
0	0.96	0.92	0.94	8215
1	0.92	0.96	0.94	8162
accuracy			0.94	16377
macro avg	0.94	0.94	0.94	16377
weighted avg	0.94	0.94	0.94	16377

- We can see most of the top predictors are from the action phase, as the drop in engagement is prominent in that phase
- Some of the factors we noticed while performing EDA which can be clubbed with these insights are:

i) Users whose maximum recharge amount is less than 200 even in the good phase, should have a tag and re-evaluated time to time as they are more likely to churn

ii) Users that have been with the network less than 4 years, should be monitored time to time, as from data we can see that users who have been associated with the network for less than 4 years tend to churn more.

- MOU is one of the major factors, but data especially VBC if the user is not using a data pack if another factor to look out

Conclusion

1. Telecom company needs to pay attention to the roaming rates. They need to provide good offers to the customers who are using services from a roaming zone.
2. The company needs to focus on the STD and ISD rates. Perhaps, the rates are too high. Provide them with some kind of STD and ISD packages.
3. To look into both of the issues stated above, it is desired that the telecom company collects customer query and complaint data and work on their services according to the needs of customers.