Telecom Churn Case Study

By
Souradeep Bose
Sai Kumar
Subhajit Bera

Business problem overview

In the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another.

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.

Definitions of Churn

Revenue-based churn:

Customers who have not utilised any revenue-generating facilities such as mobile internet, outgoing calls, SMS etc. over a given period of time. One could also use aggregate metrics such as 'customers who have generated less than INR 4 per month in total/average/median revenue'.

The main shortcoming of this definition is that there are customers who only receive calls/SMSes from their wage-earning counterparts, i.e. they don't generate revenue but use the services. For example, many users in rural areas only receive calls from their wage-earning siblings in urban areas.

Usage-based churn:

Customers who have not done any usage, either incoming or outgoing - in terms of calls, internet etc. over a period of time.

A potential shortcoming of this definition is that when the customer has stopped using the services for a while, it may be too late to take any corrective actions to retain them. For e.g., if you define churn based on a 'two-months zero usage' period, predicting churn could be useless since by that time the customer would have already switched to another operator.

Understanding the business objective and data

The dataset contains customer-level information for a span of four consecutive months - June, July, August and September. The months are encoded as 6, 7, 8 and 9, respectively.

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

Customer Behaviour During Churn

Good Phase:

In this phase, the customer is happy with the service and behaves as usual.

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.

Churn phase:

In this phase, the customer is said to have churned.

You 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 you for prediction. Thus, after tagging churn as 1/0 based on this phase, you discard all data corresponding to this phase.

Strategy

- ❑ Import data for Reading and Understanding
- ❑ Data Cleaning and Preparation for further analysis

a) Handling Missing Values

b) Outlier Treatment

c) Derive New Features

- ❑ Exploratory Data Analysis

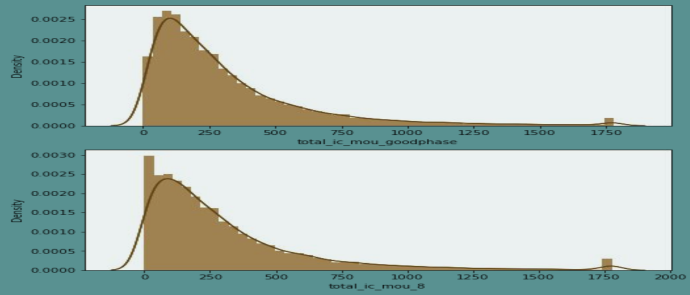
a) Univariate Analysis

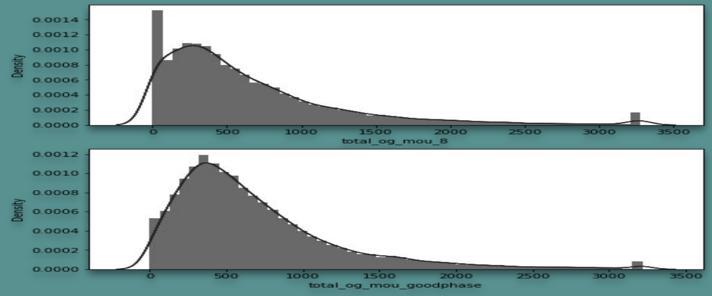
b) Bivariate Analysis

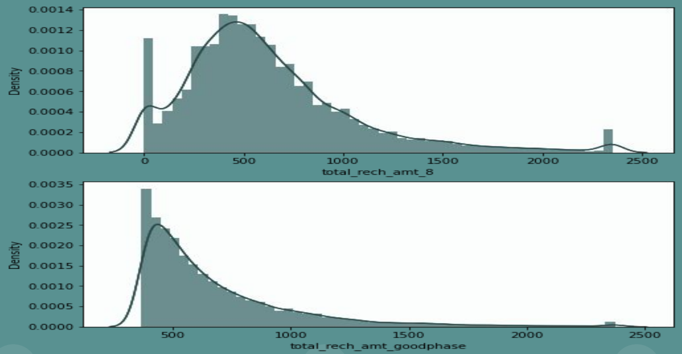
- ❑ Train Test split of data
- ❑ Performing Oversampling with SMOTE
- ❑ Feature Scaling
- ❑ Model Building
- ❑ Feature Importance and Model Interpretation
- ❑ Conclusion

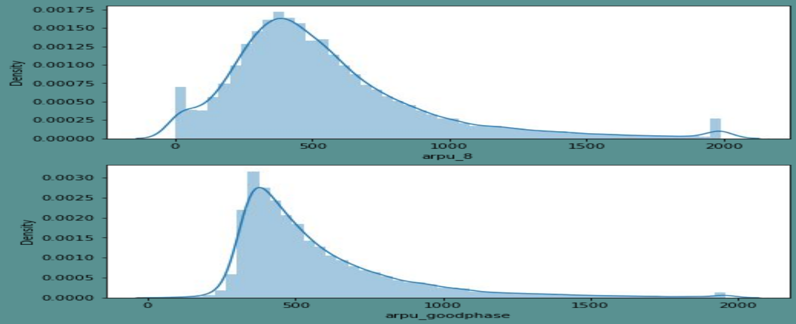
EDA (EXPLORATORY DATA ANALYSIS)

Univariate analysis





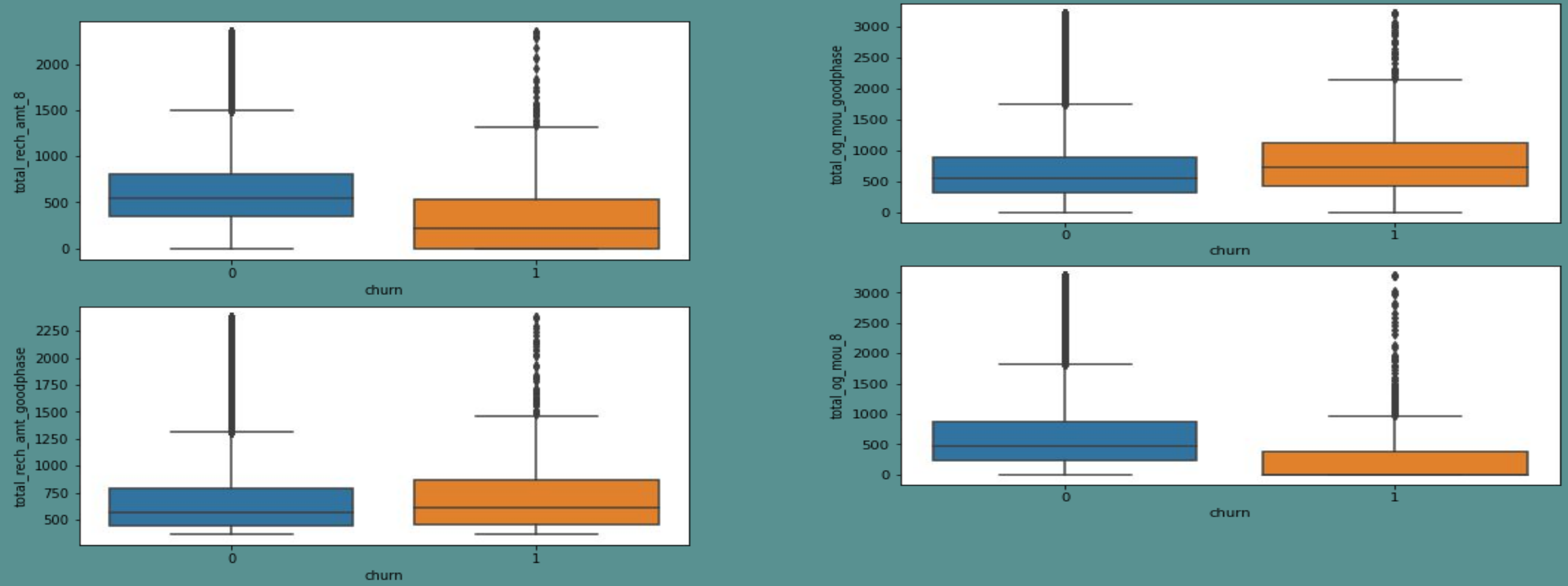




The above plots show that during the 8th month or action phase the distribution shows more density around 0 and the distribution getting more right-skewed

This can be interpreted as that during the action more users decrease there usage and recharge which is a good indicator of churn

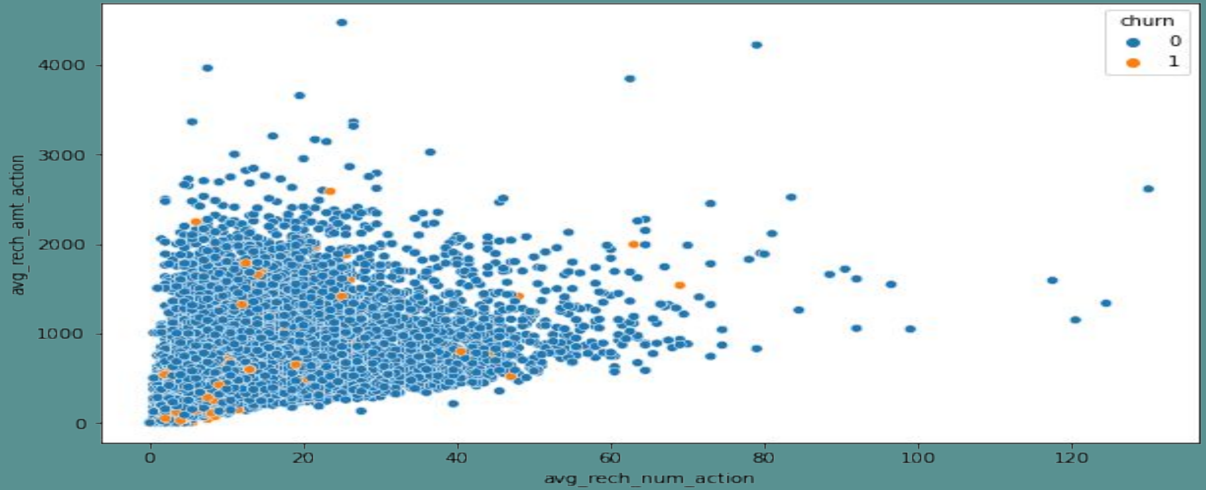
Bivariate analysis



The above boxplots clearly depict that during the action phase the users who are about to churn have lower usage in terms of phone calls and they recharge less

Bivariate analysis

Analysis of recharge amount and number of recharge in action month



We can see from the above pattern that the recharge number and the recharge amount are mostly propotional. More the number of recharge, more the amount of the recharge.

Building the first regression model

Dep. Variable:	churn	**No. Observations:**	37838
Model:	GLM	**Df Residuals:**	37817
Model Family:	Binomial	**Df Model:**	20
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-12758.
date:	Sun, 01 Oct 2023	**Deviance:**	25517.
Time:	14:38:52	**Pearson chi2:**	1.38e+05
No. Iterations:	14		
Covariance Type:	nonrobust		

Building the first regression mode

Generalized Linear Model Regression Results

Dep. Variable:	churn	**No. Observations:**	37838
Model:	GLM	**Df Residuals:**	37817
Model Family:	Binomial	**Df Model:**	20
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-12758.
Date:	Sun, 01 Oct 2023	**Deviance:**	25517.
Time:	14:38:52	**Pearson chi2:**	1.38e+05
No. Iterations:	14		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.4785	0.046	32.016	0.000	1.388	1.569
onnet_mou_8	-8.8355	0.381	-23.169	0.000	-9.583	-8.088
offnet_mou_8	-8.6665	0.375	-23.116	0.000	-9.401	-7.932
roam_og_mou_8	6.6910	0.153	43.714	0.000	6.391	6.991
std_og_mou_8	9.0810	0.454	19.993	0.000	8.191	9.971
loc_ic_t2t_mou_8	15.4212	19.814	0.778	0.436	-23.414	54.257
loc_ic_t2m_mou_8	26.3895	30.849	0.855	0.392	-34.074	86.853
loc_ic_t2f_mou_8	1.9226	5.989	0.321	0.748	-9.816	13.662
loc_ic_mou_8	-45.7042	45.850	-0.997	0.319	-135.568	44.159
std_ic_t2t_mou_8	-2.3582	0.215	-10.993	0.000	-2.779	-1.938
spl_ic_mou_8	-2.4866	0.159	-15.636	0.000	-2.798	-2.175
ic_others_8	-2.6050	0.290	-8.991	0.000	-3.173	-2.037
total_rech_num_8	-2.2900	0.129	-17.807	0.000	-2.542	-2.038
last_day_rch_amt_8	-3.4908	0.144	-24.273	0.000	-3.773	-3.209
vol_2g_mb_8	-3.6871	0.233	-15.835	0.000	-4.143	-3.231
aug_vbc_3g	-3.6998	0.242	-15.317	0.000	-4.173	-3.226
onnet_mou_goodphase	2.4275	0.127	19.054	0.000	2.178	2.677
offnet_mou_goodphase	2.6394	0.138	19.083	0.000	2.368	2.911
loc_og_t2m_mou_goodphase	-3.1336	0.236	-13.302	0.000	-3.595	-2.672
std_ic_t2t_mou_goodphase	1.9888	0.184	10.814	0.000	1.628	2.349
last_day_rch_amt_goodphase	-1.9592	0.182	-10.743	0.000	-2.317	-1.602

Dropping loc_ic_mou_8 based on high VIF

	Features	VIF
7	loc_ic_mou_8	57.91
5	loc_ic_t2m_mou_8	30.35
3	std_og_mou_8	11.66
4	loc_ic_t2t_mou_8	10.42
1	offnet_mou_8	8.57
0	onnet_mou_8	7.03
17	loc_og_t2m_mou_goodphase	3.88
16	offnet_mou_goodphase	3.49
19	last_day_rch_amt_goodphase	2.87
11	total_rech_num_8	2.81
15	onnet_mou_goodphase	2.72
18	std_ic_t2t_mou_goodphase	2.20
12	last_day_rch_amt_8	2.19
8	std_ic_t2t_mou_8	2.08
6	loc_ic_t2f_mou_8	2.04
2	roam_og_mou_8	1.82
14	aug_vbc_3g	1.23
13	vol_2g_mb_8	1.12
10	ic_others_8	1.09
9	spl_ic_mou_8	1.08

Building the second model

Dep. Variable:	churn	**No. Observations:**	37838
Model:	GLM	**Df Residuals:**	37818
Model Family:	Binomial	**Df Model:**	19
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-12759.
Date:	Sun, 01 Oct 2023	**Deviance:**	25518.
Time:	14:39:05	**Pearson chi2:**	1.38e+05
No. Iterations:	7		
Covariance Type:	nonrobust		

Generalized Linear Model Regression Results

Dep. Variable:	churn	**No. Observations:**	37838
Model:	GLM	**Df Residuals:**	37818
Model Family:	Binomial	**Df Model:**	19
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-12759.
Date:	Sun, 01 Oct 2023	**Deviance:**	25518.
Time:	14:39:05	**Pearson chi2:**	1.38e+05
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.4789	0.046	32.030	0.000	1.388	1.569
onnet_mou_8	-8.8251	0.381	-23.167	0.000	-9.572	-8.078
offnet_mou_8	-8.6554	0.374	-23.116	0.000	-9.389	-7.922
roam_og_mou_8	6.6883	0.153	43.731	0.000	6.389	6.988
std_og_mou_8	9.0699	0.454	19.987	0.000	8.181	9.959
loc_ic_t2t_mou_8	-4.3743	0.352	-12.432	0.000	-5.064	-3.685
loc_ic_t2m_mou_8	-4.3808	0.340	-12.874	0.000	-5.048	-3.714
loc_ic_t2f_mou_8	-4.0454	0.352	-11.482	0.000	-4.736	-3.355
std_ic_t2t_mou_8	-2.3562	0.214	-10.988	0.000	-2.777	-1.936
spl_ic_mou_8	-2.4868	0.159	-15.636	0.000	-2.798	-2.175
ic_others_8	-2.6038	0.290	-8.987	0.000	-3.172	-2.036
total_rech_num_8	-2.2898	0.129	-17.806	0.000	-2.542	-2.038
last_day_rch_amt_8	-3.4906	0.144	-24.270	0.000	-3.772	-3.209
vol_2g_mb_8	-3.6870	0.233	-15.835	0.000	-4.143	-3.231
aug_vbc_3g	-3.6997	0.242	-15.316	0.000	-4.173	-3.226
onnet_mou_goodphase	2.4274	0.127	19.053	0.000	2.178	2.677
offnet_mou_goodphase	2.6383	0.138	19.078	0.000	2.367	2.909
loc_og_t2m_mou_goodphase	-3.1309	0.235	-13.299	0.000	-3.592	-2.669
std_ic_t2t_mou_goodphase	1.9865	0.184	10.808	0.000	1.626	2.347
last_day_rch_amt_goodphase	-1.9587	0.182	-10.739	0.000	-2.316	-1.601

	Features	VIF
3	std_og_mou_8	11.65
1	offnet_mou_8	8.55
0	onnet_mou_8	7.02
16	loc_og_t2m_mou_goodphase	3.86
15	offnet_mou_goodphase	3.49
5	loc_ic_t2m_mou_8	2.89
18	last_day_rch_amt_goodphase	2.87
10	total_rech_num_8	2.81
14	onnet_mou_goodphase	2.72
17	std_ic_t2t_mou_goodphase	2.20
11	last_day_rch_amt_8	2.19
7	std_ic_t2t_mou_8	2.08
2	roam_og_mou_8	1.82
4	loc_ic_t2t_mou_8	1.74
6	loc_ic_t2f_mou_8	1.39
13	aug_vbc_3g	1.23
12	vol_2g_mb_8	1.12
8	spl_ic_mou_8	1.08
9	ic_others_8	1.08

Dropping std_og_mou_8 based on high VIF

Building the third model

Dep. Variable:	churn	**No. Observations:**	37838
Model:	GLM	**Df Residuals:**	37819
Model Family:	Binomial	**Df Model:**	18
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-13000.
Date:	Sun, 01 Oct 2023	**Deviance:**	26001.
Time:	14:39:14	**Pearson chi2:**	1.68e+05
No. Iterations:	7		
Covariance Type:	nonrobust		

Generalized Linear Model Regression Results

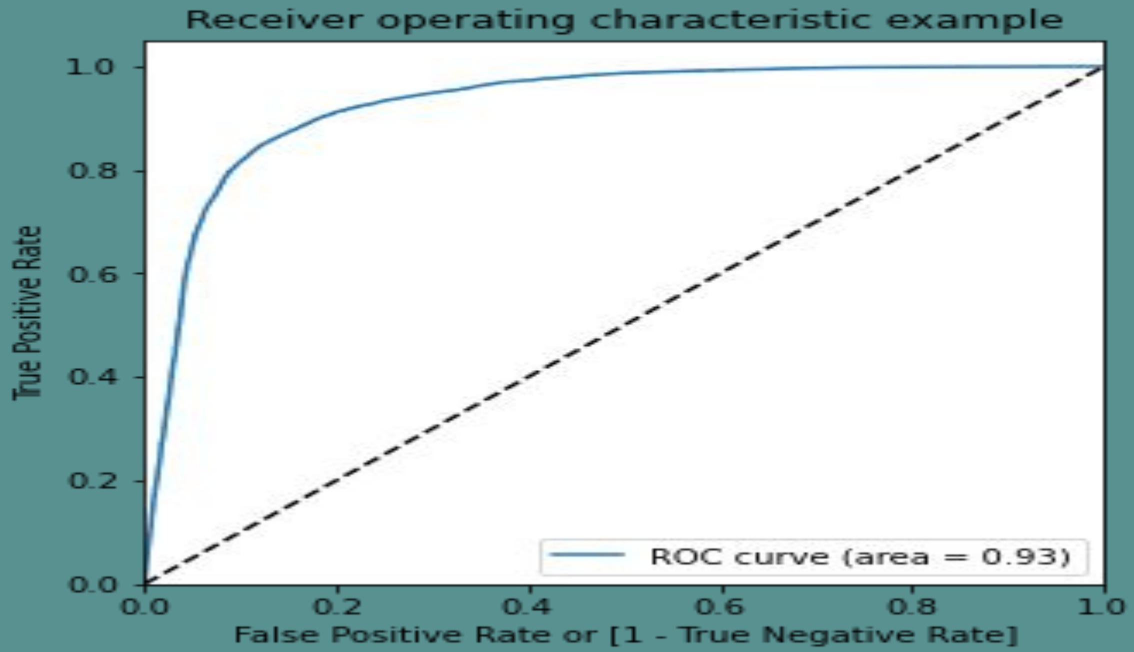
Dep. Variable:	churn	**No. Observations:**	37838
Model:	GLM	**Df Residuals:**	37819
Model Family:	Binomial	**Df Model:**	18
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-13000.
Date:	Sun, 01 Oct 2023	**Deviance:**	26001.
Time:	14:39:14	**Pearson chi2:**	1.68e+05
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.4923	0.046	32.480	0.000	1.402	1.582
onnet_mou_8	-1.8845	0.143	-13.216	0.000	-2.164	-1.605
offnet_mou_8	-1.8973	0.149	-12.694	0.000	-2.190	-1.604
roam_og_mou_8	4.9000	0.117	42.012	0.000	4.671	5.129
loc_ic_t2t_mou_8	-5.9399	0.354	-16.777	0.000	-6.634	-5.246
loc_ic_t2m_mou_8	-5.9590	0.341	-17.479	0.000	-6.627	-5.291
loc_ic_t2f_mou_8	-4.2570	0.352	-12.084	0.000	-4.948	-3.567
std_ic_t2t_mou_8	-2.3550	0.214	-11.013	0.000	-2.774	-1.936
spl_ic_mou_8	-2.5601	0.159	-16.079	0.000	-2.872	-2.248
ic_others_8	-2.6760	0.292	-9.162	0.000	-3.248	-2.104
total_rech_num_8	-2.2237	0.127	-17.551	0.000	-2.472	-1.975
last_day_rch_amt_8	-3.5100	0.143	-24.488	0.000	-3.791	-3.229
vol_2g_mb_8	-3.7060	0.233	-15.882	0.000	-4.163	-3.249
aug_vbc_3g	-3.6699	0.240	-15.297	0.000	-4.140	-3.200
onnet_mou_goodphase	2.5081	0.127	19.825	0.000	2.260	2.756
offnet_mou_goodphase	3.1360	0.139	22.634	0.000	2.864	3.408
loc_og_t2m_mou_goodphase	-4.9212	0.214	-22.965	0.000	-5.341	-4.501
std_ic_t2t_mou_goodphase	2.1131	0.183	11.571	0.000	1.755	2.471
last_day_rch_amt_goodphase	-2.2428	0.179	-12.499	0.000	-2.595	-1.891

	Features	VIF
1	offnet_mou_8	3.36
14	offnet_mou_goodphase	3.33
17	last_day_rch_amt_goodphase	2.87
4	loc_ic_t2m_mou_8	2.86
15	loc_og_t2m_mou_goodphase	2.80
9	total_rech_num_8	2.79
13	onnet_mou_goodphase	2.71
0	onnet_mou_8	2.71
16	std_ic_t2t_mou_goodphase	2.20
10	last_day_rch_amt_8	2.19
6	std_ic_t2t_mou_8	2.08
3	loc_ic_t2t_mou_8	1.63
2	roam_og_mou_8	1.40
5	loc_ic_t2f_mou_8	1.39
12	aug_vbc_3g	1.22
11	vol_2g_mb_8	1.12
7	spl_ic_mou_8	1.08
8	ic_others_8	1.08

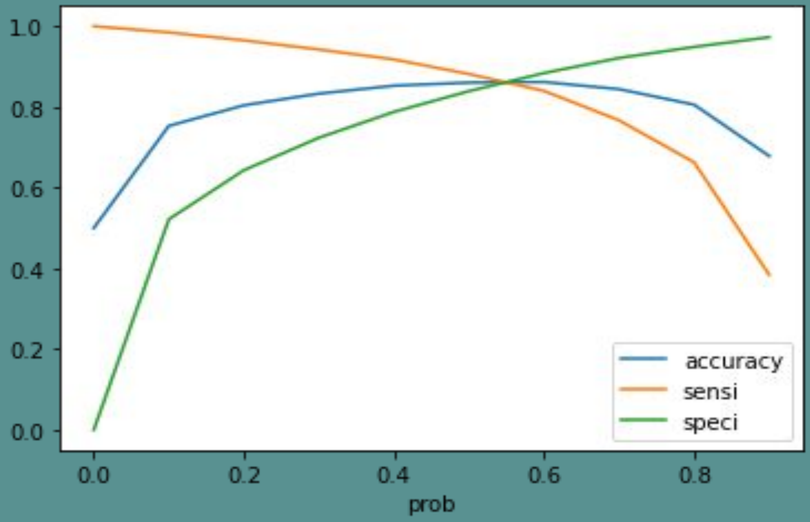
All the VIFs are below 5 and therefore we can consider this model to be the final one

Plotting the ROC curve

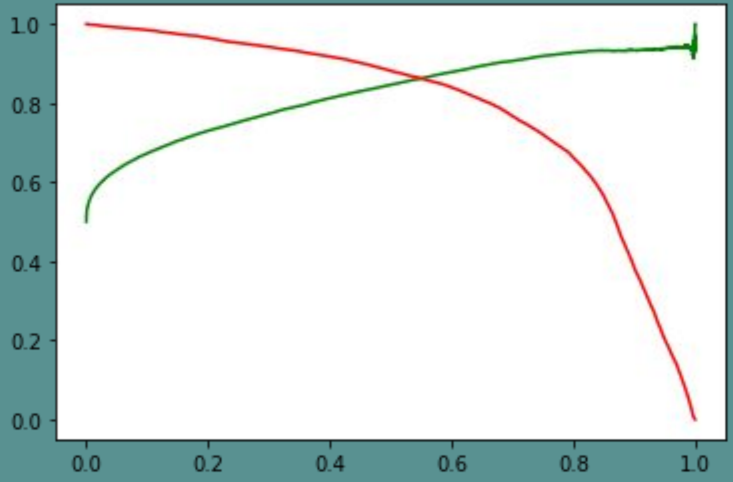


The AUC for the ROC curve is 0.93 which shows that the model is significant

Plotting accuracy sensitivity and specificity for various probabilities



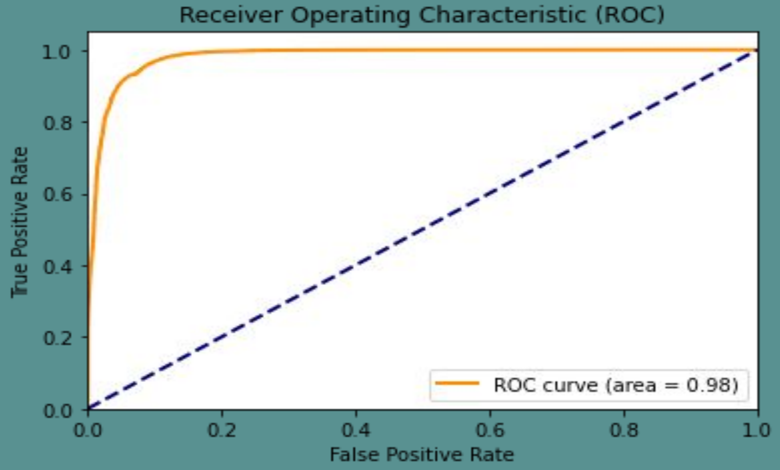
Precision and recall curve



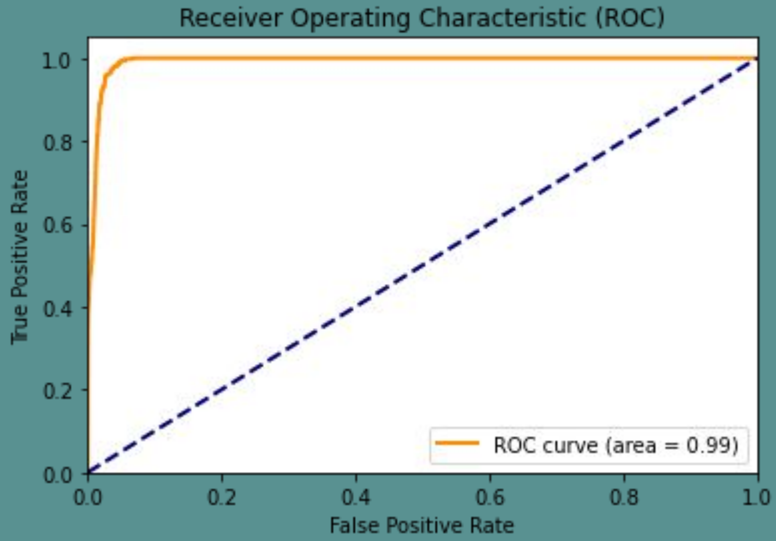
Based on the precision and recall curve and the accuracy, sensitivity and specificiy plot, we can assume 0.5 as the optimal cutoff point

Building Random forest model

Estimating the best accuracy score:0.968



We see a ROC curve of 0.98 which shows that the random forest is performing better than the logistic regression



We see a ROC curve of 0.99 which shows that the random forest is performing better than the logistic regression

BEFORE HYPERPARAMETER TUNING

After HYPERPARAMETER TUNING

Evaluating the model: Random Forest

Parameters	Train Data	Test Data
Accuracy	98%	97%
Sensitivity	43%	26%
Specificity	99.7%	99.16%

Evaluating the model: Logistic Regression

Parameters	Train Data	Test Data
Accuracy	86%	85%
Sensitivity	88%	85%
Specificity	84%	83%

Identifying important features

Variables	Coefficient	Variables	Coefficient
roam_og_mou_8	4.899963	std_ic_t2t_mou_8	-2.354985
offnet_mou_goodphase	3.136045	spl_ic_mou_8	-2.5601
onnet_mou_goodphase	2.508063	ic_others_8	-2.676003
std_ic_t2t_mou_goodphase	2.113109	last_day_rch_amt_8	-3.509998
const	1.492333	aug_vbc_3g	-3.669895
onnet_mou_8	-1.884509	vol_2g_mb_8	-3.706046
offnet_mou_8	-1.897319	loc_ic_t2f_mou_8	-4.257045
total_rech_num_8	-2.223718	loc_og_t2m_mou_goodphase	-4.921198
last_day_rch_amt_goodphase	-2.242849	loc_ic_t2t_mou_8	-5.939872
loc_ic_t2m_mou_8	-5.958982		

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_t2m_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.

Recommendations to predict the churn customers and for better business:

1. Target the customers, whose minutes of usage of the incoming local calls and outgoing local calls decrease in action phase vas good phase
2. Target the customers, with higher recharge amount in good phase.
3. Customers with higher roaming outgoing in the action phase are more likely to churn, therefore the company should provide good plans to customers moving to a different location and on a roaming plan
4. Customers whose volume based cost decreases during action phase are more likely to churn and should be targeted
5. CUSTOMERS whose local incoming call minutes of usage with other operator mobile decreases and STD incoming minutes of usage with the same operator decreases in the month of august most likely to get churned, targeting such customer with attractive local and STD offers decrease the chances of churning..
6. If the customers last day recharge amount in good phase and in the month of august decreases are more likely to be churned companies should focus and provide customers long term plans with exclusive benefits likely decrease chances of churning
7. Special incoming call minutes of usage decreases in the month of August more likely to be churned
8. Providing customization option to the customers to choose the pack based on their personal usage pattern,preferences and place of stay ,will likely decrease the chances of churning.

9. For the customers classified as a probable churn, provide the customers with attractive offers they cannot resist and retain them.
10. Provide offers on long term plans so that the customer would be loyal.
11. Provide the customers offers based on their usage and profile.

THANK YOU.