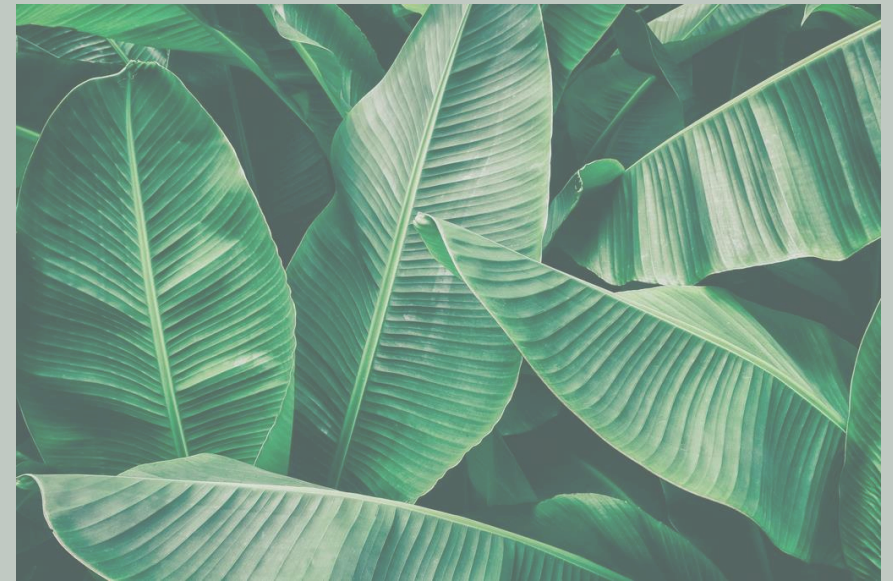


---

# Telecom Churn Case Study



By,

Krishna Prasad Ponnur,

Kowshik Sarma,

Koripadu Lavanya

# Agenda

01 Business Problem Statement

02 Steps Involved

03 Data cleaning

03 EDA

- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis

04 Class Imbalance

05 Model Selection

06 Model Building

07 Model Evaluation

08 Important Indicators

09 Insights

# Business Problem Statement

- ❑ In the telecom industry, 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.
- ❑ To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.
- ❑ Approximately 80% of revenue comes from the top 20% of customers (called high-value customers). Thus, if we can reduce the churn of high-value customers, we will be able to reduce significant revenue leakage.

*Retaining high profitable customers is the number one business goal.*

# Steps Involved

Missing value imputation

Exploratory Data Analysis

Model Selection :

To find the best model that could forecast consumer behavior, extensive experimentation with a larger number of models was conducted. This allowed the business to take proactive measures to keep the clients.

Models tried to give the best are:

- Logistic regression with RFE and manual elimination,
- Decision tree,
- Random forests

with Random under-sampling, Tomek Links, Random over-sampling, SMOTE, ADASYN and SMOTE+Tomek.

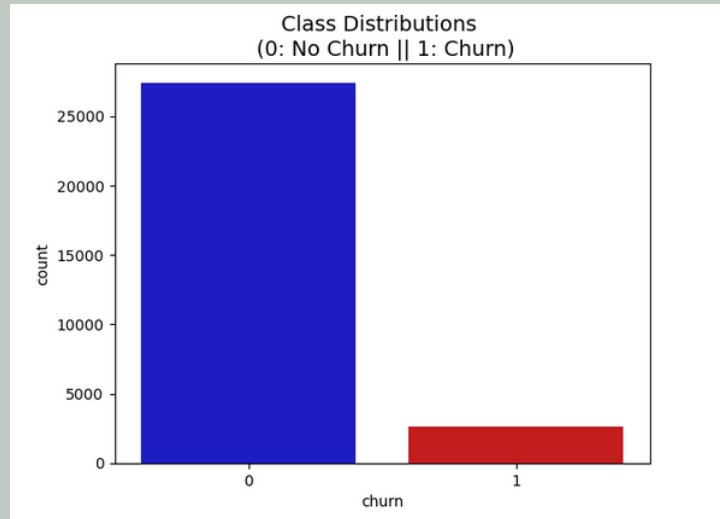
Model Building

Model Evaluation

Insights

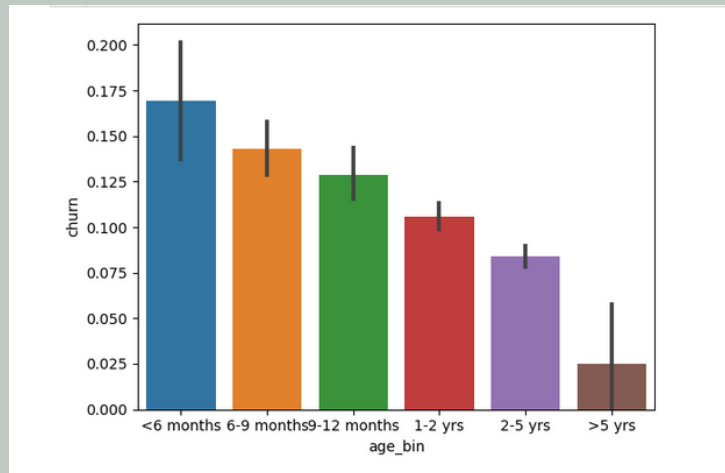


# Exploratory Data Analysis - Univariate Analysis



There is high class imbalance and , it has been handled.

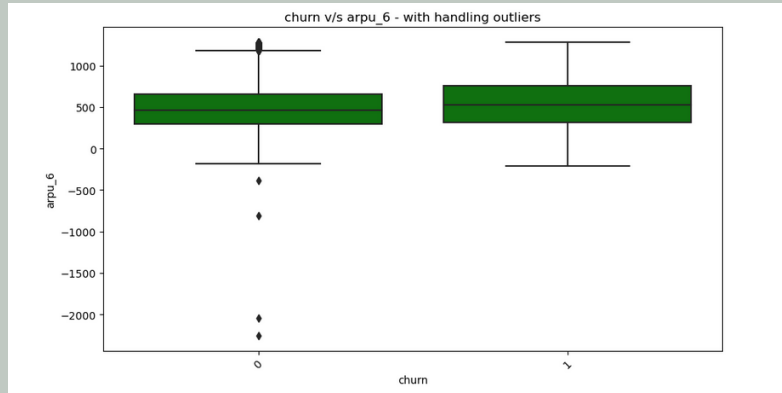
Churn



The churn rate rises in comparison to new consumers. After utilizing the network for a while, users are generally satisfied.

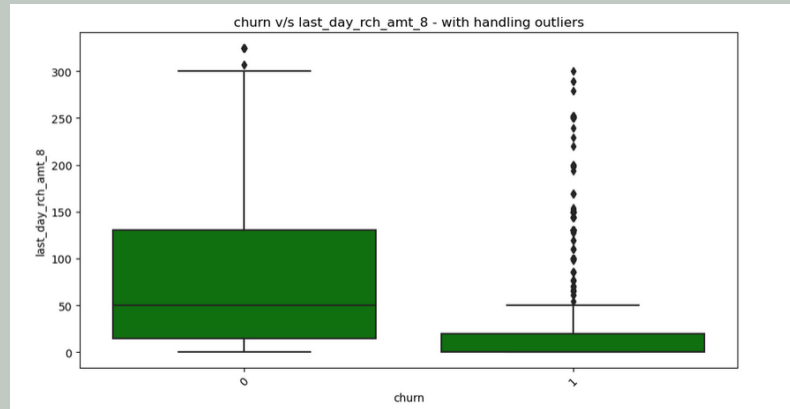
aon\_month

# Exploratory Data Analysis - Bivariate Analysis



ARPU\_6

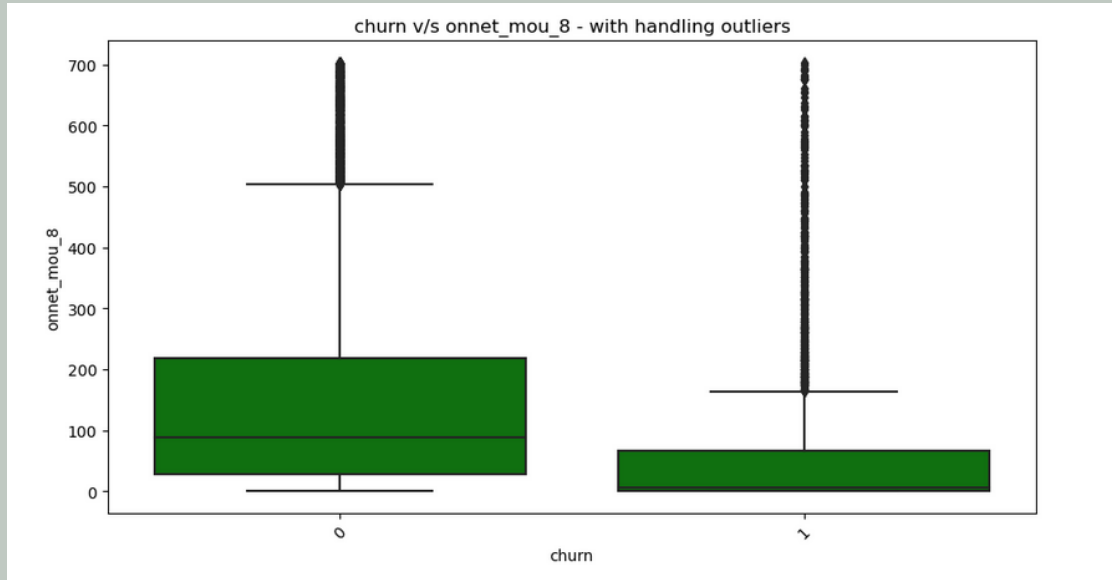
The Minimum average revenue per user (ARPU) lies between 250 to 500.



last\_day\_rch\_amt\_8

The median last day recharge amount for month 8, during the action phase, is almost zero, indicating that churned customers did not recharge at the end of this month. Consequently, it's apparent that their purchasing activity significantly decreases in month 8, which precedes their churn in month 9.

# Exploratory Data Analysis - Bivariate Analysis



onnet\_mou\_8

The median on-network minutes of usage for month 8, which corresponds to the action phase, is significantly lower for churned customers when compared to non-churned customers. This observation strongly indicates a decreasing pattern in their network usage leading up to their churn.

# Exploratory Data Analysis - Correlation Analysis

```
['isd_og_mou_7',  
 'isd_og_mou_8',  
 'total_rech_amt_6',  
 'total_rech_amt_7',  
 'total_rech_amt_8']
```

Correlated variables

As the data set contains numerous variables, its difficult to deduce a heat map for all. So, we have checked correlation matrix for all variables and derived these highly correlated variables. Eliminated these highly correlated variables.



# Class Imbalance Handling

- Class imbalance can significantly impact model performance. To address this issue, several techniques, including under sampling, oversampling, SMOTE (Synthetic Minority Over-sampling Technique), Tomek links , ADASYN and SMOTE+TOMEK have been employed in our model.
- The above-mentioned techniques are performed on Logistic Regression , Decision tree and Random forest Models.



# Model Selection

Both Logistic Regression and Random Forest have high Accuracy, Precision, Recall, and F1 scores for the SMOTE. Using RFE and manual exclusion, we will construct a logistic model on the SMOTE resampling.

	Logistic Regression				Decision Tree				Random Forest			
	Accuracy	Precision	Recall	F1 score	Accuracy	Precision	Recall	F1 score	Accuracy	Precision	Recall	F1 score
Base	0.9204	0.709	0.134	0.226	0.90781	0.4675	0.4807	0.474	0.94	0.728	0.478	0.578
Random US	0.8005	0.278	0.819	0.414	0.7794	0.2495	0.773	0.3774	0.889	0.424	0.788	0.551
Tomek Links	0.919	0.647	0.141	0.232	0.9088	0.474	0.511	0.4922	0.941	0.729	0.501	0.594
Random OS	0.811	0.293	0.838	0.433	0.912	0.491	0.49	0.49	0.941	0.704	0.54	0.611
SMOTE	0.825	0.308	0.819	0.447	0.875	0.362	0.575	0.443	0.926	0.561	0.672	0.611
ADASYN	0.807	0.29	0.844	0.431	0.872	0.354	0.582	0.441	0.923	0.547	0.683	0.608
SMOTE+TOMEK	0.825	0.308	0.818	0.447	0.875	0.361	0.574	0.443	0.925	0.554	0.674	0.609

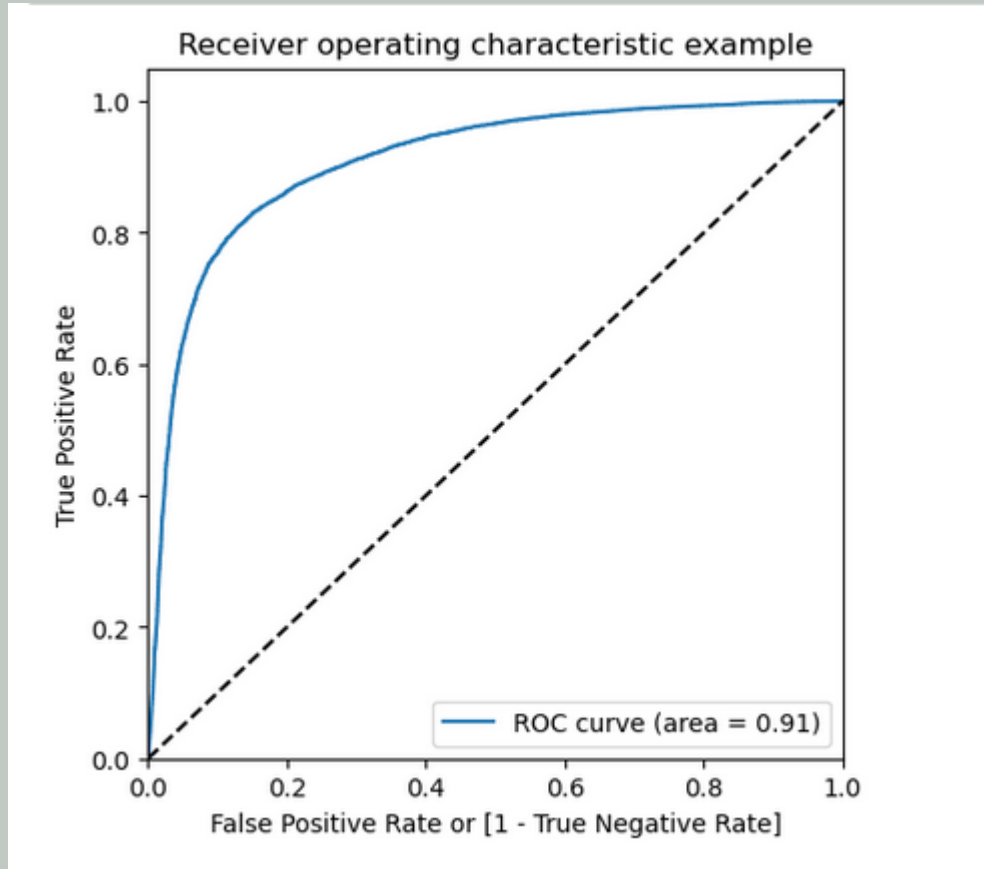
# Model 1 , 2 & 3 on Logistic Regression

	Features	VIF
0	arpu_6	5.34
4	loc_ic_t2m_mou_7	5.33
8	total_rech_num_6	3.99
6	total_ic_mou_7	3.97
5	loc_ic_t2m_mou_8	3.68
9	total_rech_num_8	2.95
2	loc_og_t2m_mou_8	2.11
3	std_og_mou_7	1.77
10	last_day_rch_amt_8	1.49
12	sachet_2g_8	1.23
1	roam_og_mou_7	1.15
13	monthly_3g_8	1.15
11	monthly_2g_8	1.10
7	spl_ic_mou_8	1.04
14	sep_vbc_3g	1.04

	coef	std err	z	P> z	[0.025	0.975]
const	0.5047	0.118	4.277	0.000	0.273	0.736
arpu_6	7.1771	0.865	8.298	0.000	5.482	8.872
roam_og_mou_7	8.4420	0.424	19.932	0.000	7.612	9.272
loc_og_t2m_mou_8	-14.8006	0.956	-15.480	0.000	-16.675	-12.927
std_og_mou_7	1.7165	0.202	8.503	0.000	1.321	2.112
loc_ic_t2m_mou_7	28.0582	1.044	26.866	0.000	26.011	30.105
loc_ic_t2m_mou_8	-77.0873	2.085	-36.972	0.000	-81.174	-73.001
total_ic_mou_7	-5.3369	0.633	-8.426	0.000	-6.578	-4.095
spl_ic_mou_8	-23.2337	1.480	-15.701	0.000	-26.134	-20.334
total_rech_num_6	6.7633	0.672	10.065	0.000	5.446	8.080
total_rech_num_8	-13.9459	0.398	-35.070	0.000	-14.725	-13.166
last_day_rch_amt_8	-22.9976	0.873	-26.329	0.000	-24.710	-21.286
monthly_2g_8	-7.7121	0.377	-20.435	0.000	-8.452	-6.972
sachet_2g_8	-9.0068	0.720	-12.507	0.000	-10.418	-7.595
monthly_3g_8	-14.5521	0.891	-16.335	0.000	-16.298	-12.806
sep_vbc_3g	-211.4270	21.246	-9.952	0.000	-253.068	-169.786

For Model – 3 , the p-value is 'zero' for all the variables and the VIF value is low. Based on these, Model – 3 is selected for model evaluation.

# Logistic Regression - Model Evaluation on Train set

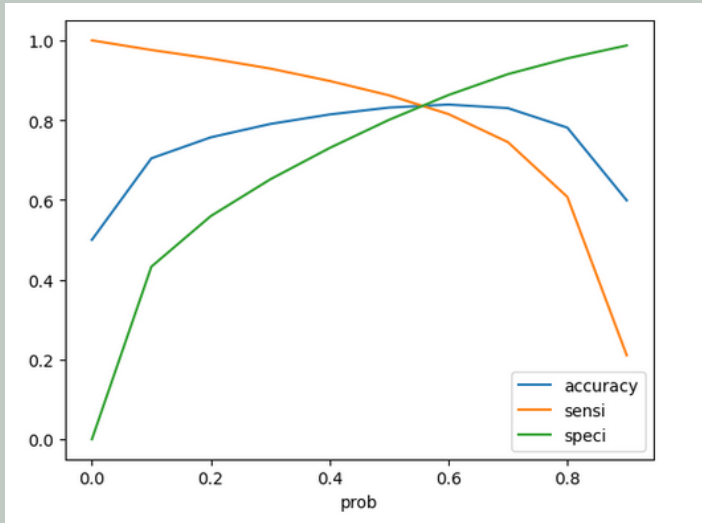


ROC curve

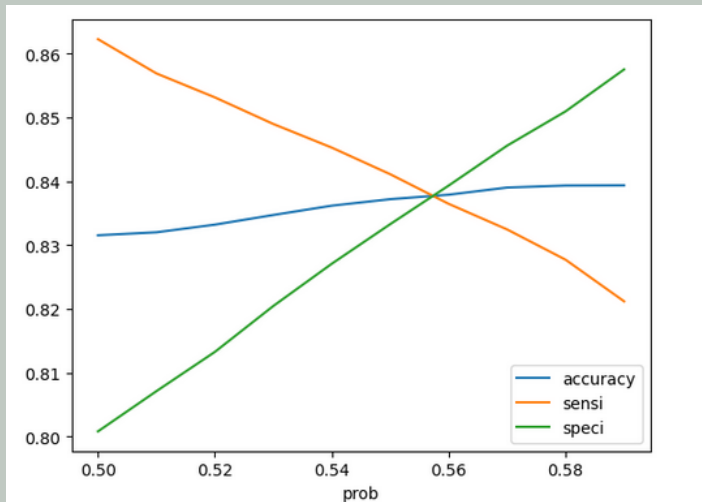
The ROC of the model is 91% which shows a strong performance.

# Logistic Regression - Model Building on Train set

Accuracy , Sensitivity and Specificity for various probability cutoffs

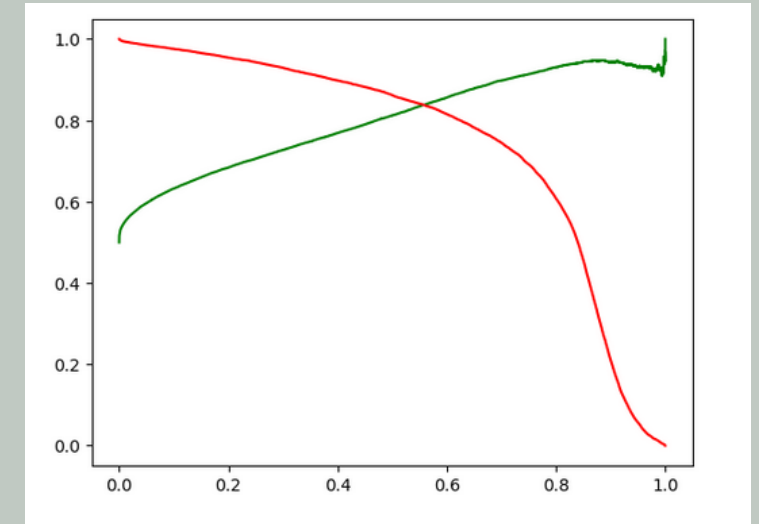


Probability cutoff



Probability cutoff between 0.5 and 0.6

The probability lies between 0.5 and 0.6



Precision , Recall curve

Intersection threshold point is 0.54

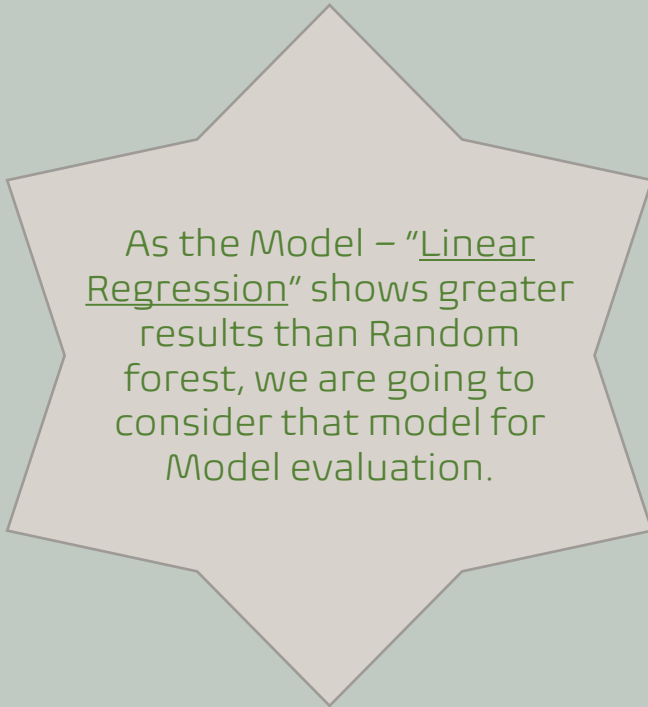
# Hyper-parameter Tuning In Random forest

Hyper-parameters like :

'max\_depth'  
'max\_features'  
'min\_samples\_leaf'  
'n\_estimators' are used.

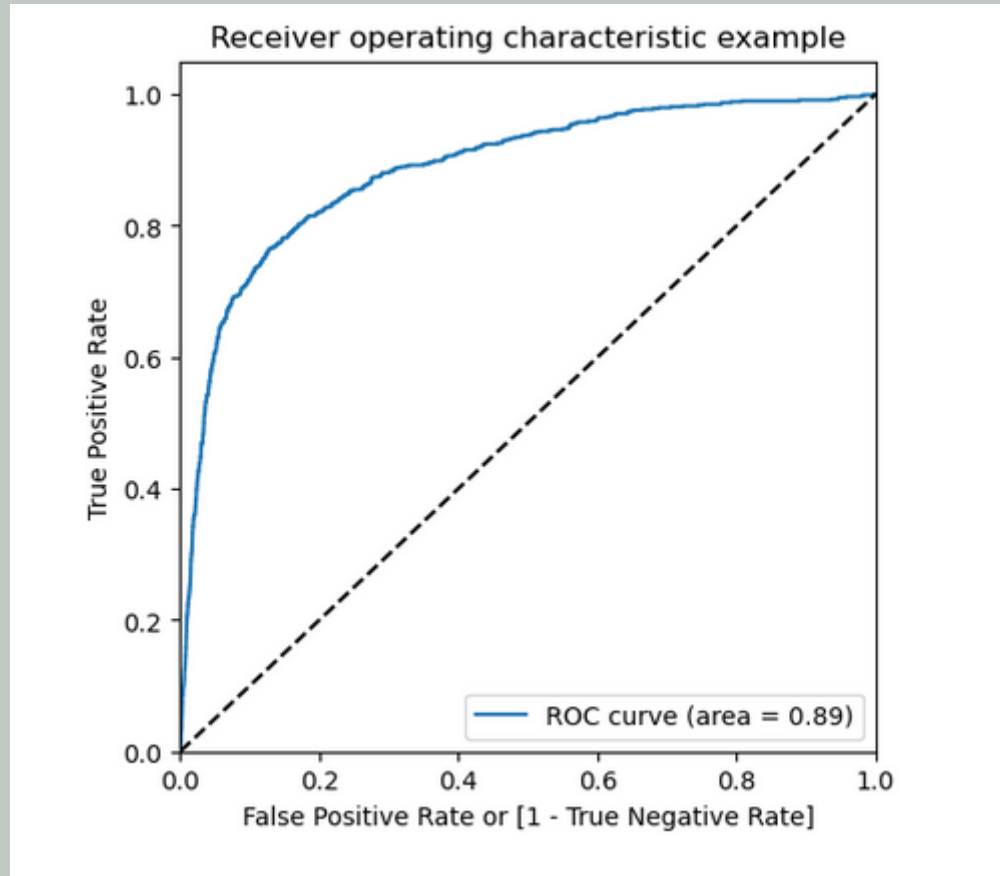
And got a 'best\_score' result as 94.2

And max\_depth=15, max\_features=3, min\_samples\_leaf=20, n\_estimators=80



As the Model – “Linear Regression” shows greater results than Random forest, we are going to consider that model for Model evaluation.

# Logistic Regression - Model Evaluation on Test set



ROC curve

The ROC of the model is 89% which shows a strong performance.

# Important indicators:

- 1.Average revenue per use for month 6
- 2.ROAM outgoing calls minutes of usage for month 7
- 3.Local outgoing (Operator T to other operator) calls minutes of usage for month 8
- 4.STD outgoing calls minutes of usage for month 7
5. Local incoming calls minutes of usage for month 7 , 8
- 6.Total incoming calls minutes of usage for month 7
7. Special incoming calls minutes of usage for month 8
- 8.Total number of recharges for the month of 6 , 8
- 9.Last day recharge amount for month 8
- 10.Mobile internet recharge for 2G in month 8
11. Sachet internet recharge for 2G in month 8
- 12.Mobile internet recharge for 3G in month 8
- 13.Volume based cost for September (month 8)

Summary of the variable based on the model:

	coef
const	0.5047
arpu_6	7.1771
roam_og_mou_7	8.4420
loc_og_t2m_mou_8	-14.8006
std_og_mou_7	1.7165
loc_ic_t2m_mou_7	28.0582
loc_ic_t2m_mou_8	-77.0873
total_ic_mou_7	-5.3369
spl_ic_mou_8	-23.2337
total_rech_num_6	6.7633
total_rech_num_8	-13.9459
last_day_rch_amt_8	-22.9976
monthly_2g_8	-7.7121
sachet_2g_8	-9.0068
monthly_3g_8	-14.5521
sep_vbc_3g	-211.4270



# Insights

- In the context of this telecom company, the primary focus is on identifying churners rather than non-churners. Customer retention takes precedence as the most critical objective. Therefore, we will select Sensitivity/Recall as our evaluation metric since it quantifies the proportion of actual positives correctly detected.
- Take note of the following essential predictor attributes, which serve as indicators of churn:
  1. The most important consumers who are prone to churn are those who have just begun utilizing the telecom network. Customers who have been utilizing the network for a longer period are content with it.
  2. The count of recharges, the recharge amount, and the last recharge date are crucial indicators of customer usage. A decrease in any of these metrics may suggest that the customer intends to wait until the remaining validity period before considering a switch to another provider. Offering discounted recharge packs during this time could help retain the customer.
  3. Internet usage is a vital variable as it signifies whether a customer is actively engaged in mobile operations like social networking, mobile banking, and bill payments. A decrease in or complete cessation of internet usage may signal a higher likelihood of churn. To retain such customers, attractive offers like bundled data packs, reduced prices, or even a month of free data can be provided to incentivize them to remain on the network.
  4. Our focus should be on, and efforts should be made to retain the clients whose average monthly revenue for the month of June is quite high.
  5. Customers that have made extremely few special calls, used very little internet use, recharged very little internet, and have used very few minutes for local outbound calls throughout the month of August likely to churn.
  6. It is essential to consistently monitor incoming STD (Standard), roaming outgoing, incoming calls and Special calls for Month 7 to detect any discernible decline in usage patterns.
  7. By examining the cost based on volume for the 9th month, we can assess whether the customer is demonstrating a commitment to remain on the network into Month 9.