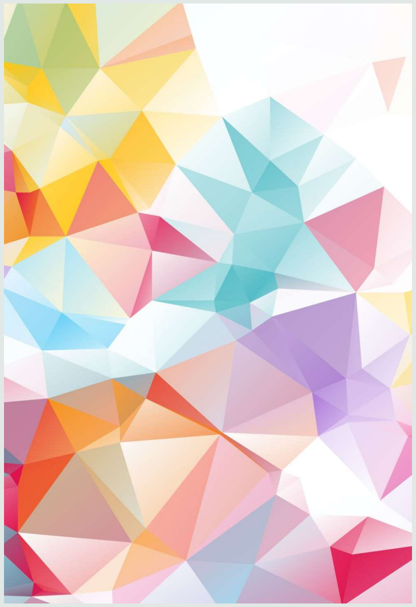
TELECOM CHURN CASE STUDY

- DS54_Course 3
- BY OMKAR NIKAM, NACHIKET PATIL & K CHETAN PAI



CASE STUDY OVERVIEW

- In the telecom industry, customers can 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.

CASE STUDY OVERVIEW

- There are two main models of payment in the telecom industry –
- Postpaid (customers pay a monthly/annual bill after using the services) and prepaid (customers pay/recharge with a certain amount in advance and then use the services). In the postpaid model, when customers want to switch to another operator, they usually inform the existing operator to terminate the services, and you directly know that this is an instance of churn.
- Prepaid - customers who want to switch to another network can simply stop using the services without any notice, and it is hard to know whether someone has churned or is simply not using the services temporarily (e.g., someone may be on a trip abroad for a month or two and then intend to resume using the services again).
- Thus, churn prediction is usually more critical (and non-trivial) for prepaid customers, and the term ‘churn’ should be defined carefully.
- Also, prepaid is the most common model in India and Southeast Asia, while postpaid is more common in Europe in North America.
- The case study is based on the Indian and Southeast Asian market.

CASE STUDY OVERVIEW

- Revenue-based churn: Customers who have not utilized any revenue-generating facilities such as mobile internet, outgoing calls, SMS etc. over a given period. 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/SMS es 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.
- 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.
- The case study focuses on the "usage-based" churn.

CASE STUDY OVERVIEW

- Customers usually do not decide to switch to another competitor instantly, but rather over a period of time (this is especially applicable to high-value customers). In churn prediction, we assume that there are three phases of the customer lifecycle :
- The ‘good’ phase: In this phase, the customer is happy with the service and behaves as usual.
- The ‘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 behavior than in the ‘good’ months. Also, it is crucial to identify high-churn-risk customers in this phase, since some corrective actions can be taken at this point (such as matching the competitor’s offer/improving the service quality etc.)
- The ‘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.
- The case study is based on four-month window, the first two months are the ‘good’ phase, the third month is the ‘action’ phase, and the fourth month is the ‘churn’ phase.

STEPS

The Main steps involved were -

➡ **1. Reading and Understanding the data.**

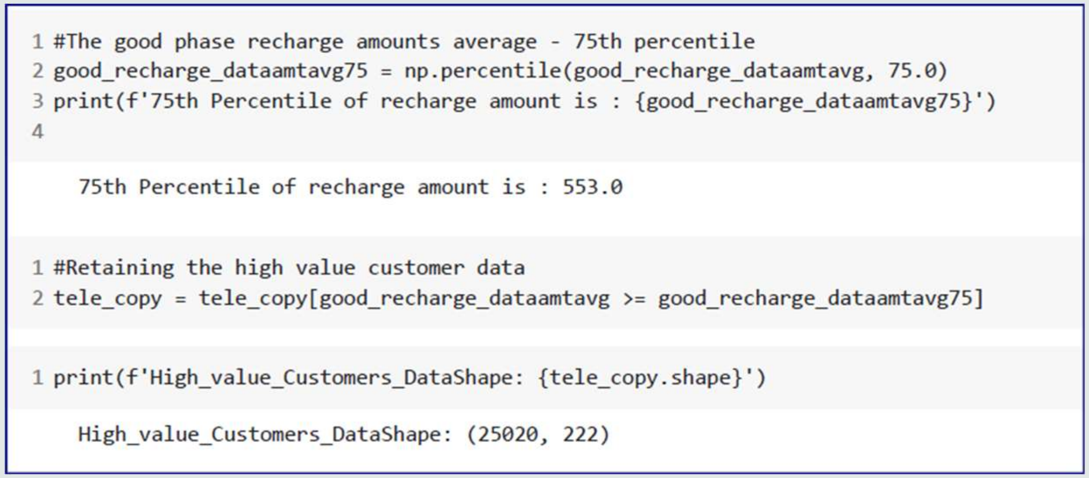
➡ **2. Finding High value customers related data.**

➡ **3. Cleaning the data, preparing the data and Exploratory data analysis.**

➡ **4. Building of the Models using Logistic Regression, Decision trees & Random Forests.**

PROCESS

- Finding the data about high value customers based on the recharge amount data.
- Finding Total recharge data amount for the 4 months.
- Retaining the High value customers data.



```
#The good phase recharge amounts average - 75th percentile
good_recharge_dataamtavg75 = np.percentile(good_recharge_dataamtavg, 75.0)
print(f'75th Percentile of recharge amount is : {good_recharge_dataamtavg75}')

```

```
75th Percentile of recharge amount is : 553.0
```

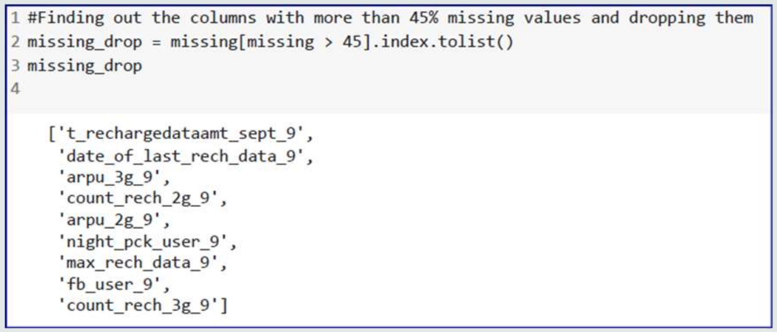
```
#Retaining the high value customer data
tele_copy = tele_copy[good_recharge_dataamtavg >= good_recharge_dataamtavg75]
```

```
print(f'High_value_Customers_DataShape: {tele_copy.shape}')
```

```
High_value_Customers_DataShape: (25020, 222)
```

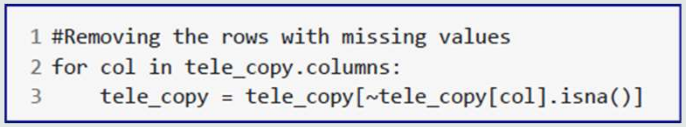
PROCESS

- Missing values columns were identified and those with more than 45% missing values were dropped and those having same values in their rows were also dropped.
- Rows with missing values were removed.

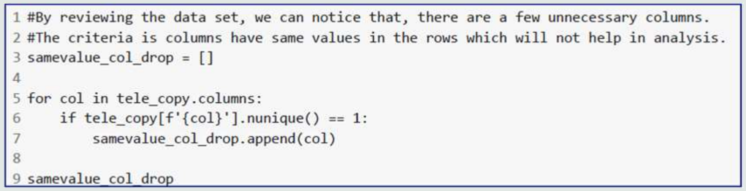


```
#Finding out the columns with more than 45% missing values and dropping them
missing_drop = missing[missing > 45].index.tolist()
missing_drop

```

```
['t_rechargedataamt_sept_9',
 'date_of_last_rech_data_9',
 'arpu_3g_9',
 'count_rech_2g_9',
 'arpu_2g_9',
 'night_pck_user_9',
 'max_rech_data_9',
 'fb_user_9',
 'count_rech_3g_9']
```



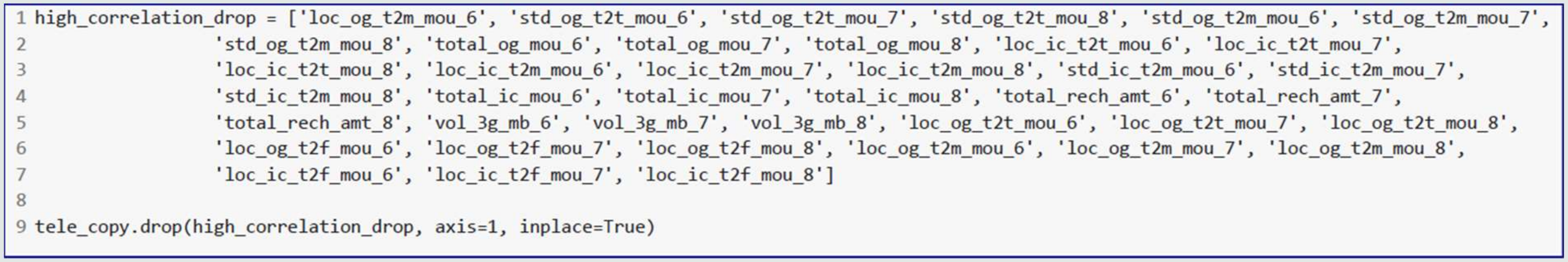
```
#Removing the rows with missing values
for col in tele_copy.columns:
    tele_copy = tele_copy[~tele_copy[col].isna()]
```



```
#By reviewing the data set, we can notice that, there are a few unnecessary columns.
#The criteria is columns have same values in the rows which will not help in analysis.
samevalue_col_drop = []

for col in tele_copy.columns:
    if tele_copy[f'{col}'].nunique() == 1:
        samevalue_col_drop.append(col)

samevalue_col_drop
```

PROCESS

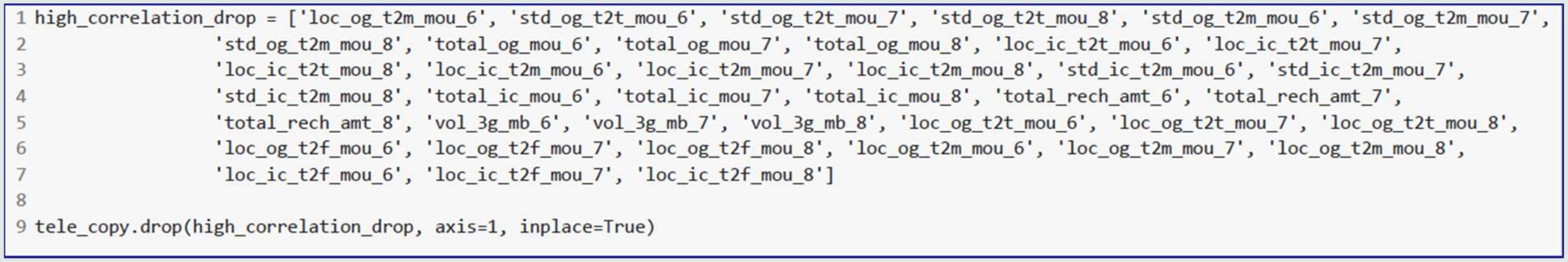
- Columns having high correlation between features were removed.



```
high_correlation_drop = ['loc_og_t2m_mou_6', 'std_og_t2t_mou_6', 'std_og_t2t_mou_7', 'std_og_t2t_mou_8', 'std_og_t2m_mou_6', 'std_og_t2m_mou_7',
                  'std_og_t2m_mou_8', 'total_og_mou_6', 'total_og_mou_7', 'total_og_mou_8', 'loc_ic_t2t_mou_6', 'loc_ic_t2t_mou_7',
                  'loc_ic_t2t_mou_8', 'loc_ic_t2m_mou_6', 'loc_ic_t2m_mou_7', 'loc_ic_t2m_mou_8', 'std_ic_t2m_mou_6', 'std_ic_t2m_mou_7',
                  'std_ic_t2m_mou_8', 'total_ic_mou_6', 'total_ic_mou_7', 'total_ic_mou_8', 'total_rech_amt_6', 'total_rech_amt_7',
                  'total_rech_amt_8', 'vol_3g_mb_6', 'vol_3g_mb_7', 'vol_3g_mb_8', 'loc_og_t2t_mou_6', 'loc_og_t2t_mou_7', 'loc_og_t2t_mou_8',
                  'loc_og_t2f_mou_6', 'loc_og_t2f_mou_7', 'loc_og_t2f_mou_8', 'loc_og_t2m_mou_6', 'loc_og_t2m_mou_7', 'loc_og_t2m_mou_8',
                  'loc_ic_t2f_mou_6', 'loc_ic_t2f_mou_7', 'loc_ic_t2f_mou_8']

tele_copy.drop(high_correlation_drop, axis=1, inplace=True)
```

PROCESS

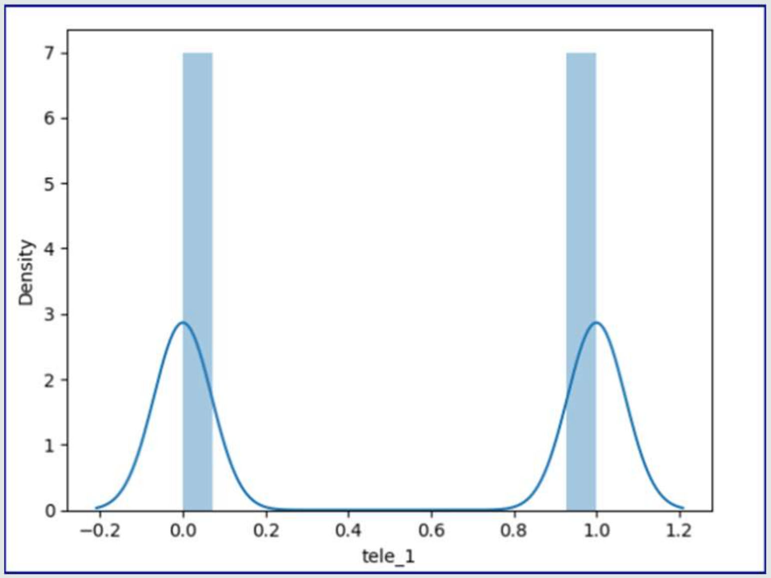
- Columns having high correlation between features were removed.



```python
high_correlation_drop = ['loc_og_t2m_mou_6', 'std_og_t2t_mou_6', 'std_og_t2t_mou_7', 'std_og_t2t_mou_8', 'std_og_t2m_mou_6', 'std_og_t2m_mou_7',
                         'std_og_t2m_mou_8', 'total_og_mou_6', 'total_og_mou_7', 'total_og_mou_8', 'loc_ic_t2t_mou_6', 'loc_ic_t2t_mou_7',
                         'loc_ic_t2t_mou_8', 'loc_ic_t2m_mou_6', 'loc_ic_t2m_mou_7', 'loc_ic_t2m_mou_8', 'std_ic_t2m_mou_6', 'std_ic_t2m_mou_7',
                         'std_ic_t2m_mou_8', 'total_ic_mou_6', 'total_ic_mou_7', 'total_ic_mou_8', 'total_rech_amt_6', 'total_rech_amt_7',
                         'total_rech_amt_8', 'vol_3g_mb_6', 'vol_3g_mb_7', 'vol_3g_mb_8', 'loc_og_t2t_mou_6', 'loc_og_t2t_mou_7', 'loc_og_t2t_mou_8',
                         'loc_og_t2f_mou_6', 'loc_og_t2f_mou_7', 'loc_og_t2f_mou_8', 'loc_og_t2m_mou_6', 'loc_og_t2m_mou_7', 'loc_og_t2m_mou_8',
                         'loc_ic_t2f_mou_6', 'loc_ic_t2f_mou_7', 'loc_ic_t2f_mou_8']

tele_copy.drop(high_correlation_drop, axis=1, inplace=True)
```

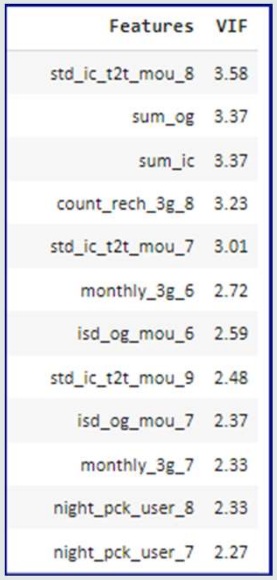
ANALYSIS

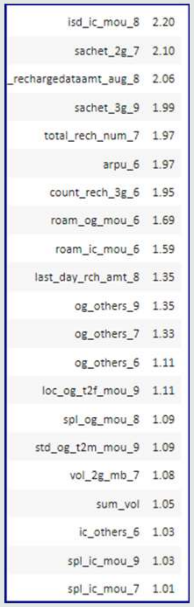
- SNS Distribution Plot –



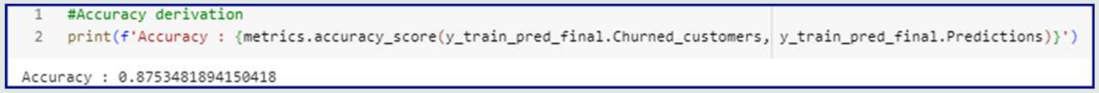
ANALYSIS

- Variance Inflation Factor and Accuracy – Model “reg_model_5” contained features with a VIF of less than 4.00 and an accuracy of 87.5% and therefore, the Model “reg_model_5” with can be considered for data analysis.
-





Features	VIF
std_ic_t2t_mou_8	3.58
sum_og	3.37
sum_ic	3.37
count_rech_3g_8	3.23
std_ic_t2t_mou_7	3.01
monthly_3g_6	2.72
isd_og_mou_6	2.59
std_ic_t2t_mou_9	2.48
isd_og_mou_7	2.37
monthly_3g_7	2.33
night_pck_user_8	2.33
night_pck_user_7	2.27
isd_ic_mou_8	2.20
sachet_2g_7	2.10
_rechargedataamt_aug_8	2.06
sachet_3g_9	1.99
total_rech_num_7	1.97
arpu_6	1.97
count_rech_3g_6	1.95
roam_og_mou_6	1.69
roam_ic_mou_6	1.59
last_day_rch_amt_8	1.35
og_others_9	1.35
og_others_7	1.33
og_others_6	1.11
loc_og_t2f_mou_9	1.11
spl_og_mou_8	1.09
std_og_t2m_mou_9	1.09
vol_2g_mb_7	1.08
sum_vol	1.05
ic_others_6	1.03
spl_ic_mou_9	1.03
spl_ic_mou_7	1.01

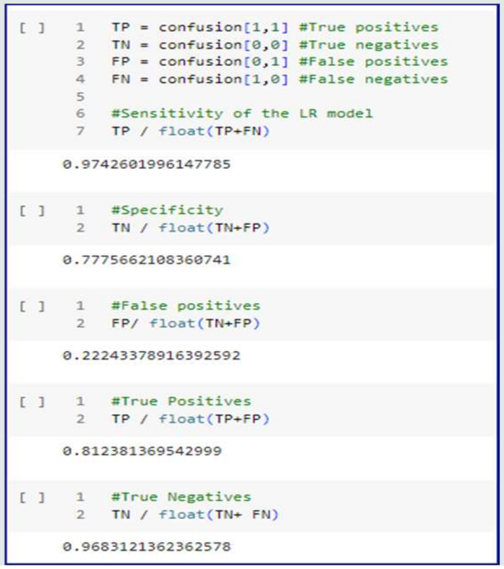


```
#Accuracy derivation
print(f'Accuracy : {metrics.accuracy_score(y_train_pred_final.Churned_customers, y_train_pred_final.Predictions)}')
```

```
Accuracy : 0.8753481894150418
```

ANALYSIS

- A **true positive** is an outcome where the model *correctly* predicts the *positive* class. Similarly, a **true negative** is an outcome where the model *correctly* predicts the *negative* class.
- A **false positive** is an outcome where the model *incorrectly* predicts the *positive* class. And a **false negative** is an outcome where the model *incorrectly* predicts the *negative* class.
- The Sensitivity of the Model is 0.97. A highly sensitive test means that there are few false negative results.
- The Specificity of the Model is 0.78.



```
TP = confusion[1,1] #True positives
TN = confusion[0,0] #True negatives
FP = confusion[0,1] #False positives
FN = confusion[1,0] #False negatives

#Sensitivity of the LR model
TP / float(TP+FN)
```

```
0.9742601996147785
```

```
#Specificity
TN / float(TN+FP)
```

```
0.7775662108360741
```

```
#False positives
FP/ float(TN+FP)
```

```
0.22243378916392592
```

```
#True Positives
TP / float(TP+FP)
```

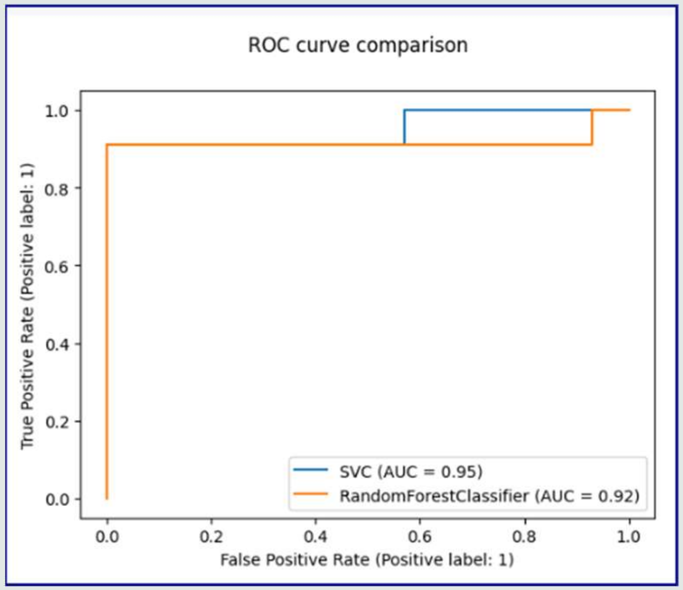
```
0.812381369542999
```

```
#True Negatives
TN / float(TN+ FN)
```

```
0.9683121362362578
```

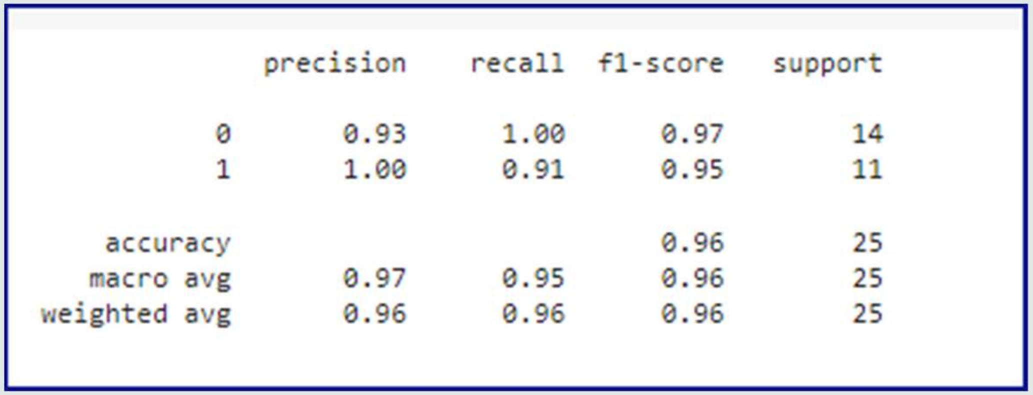
ANALYSIS

- ROC Curve comparison –



ANALYSIS

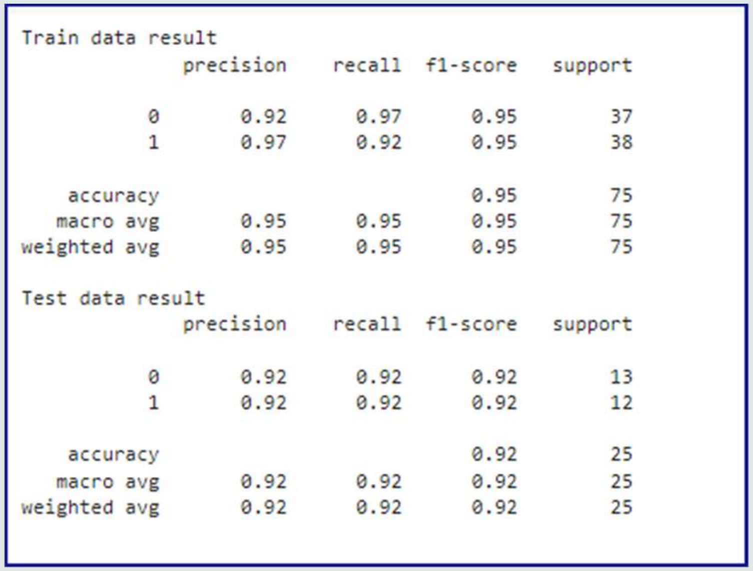
- Decision tree – The Model’s accuracy is 96%.



```
              precision    recall  f1-score   support

           0       0.93      1.00      0.97        14
           1       1.00      0.91      0.95        11

    accuracy                           0.96        25
   macro avg       0.97      0.95      0.96        25
weighted avg       0.96      0.96      0.96        25
```

ANALYSIS

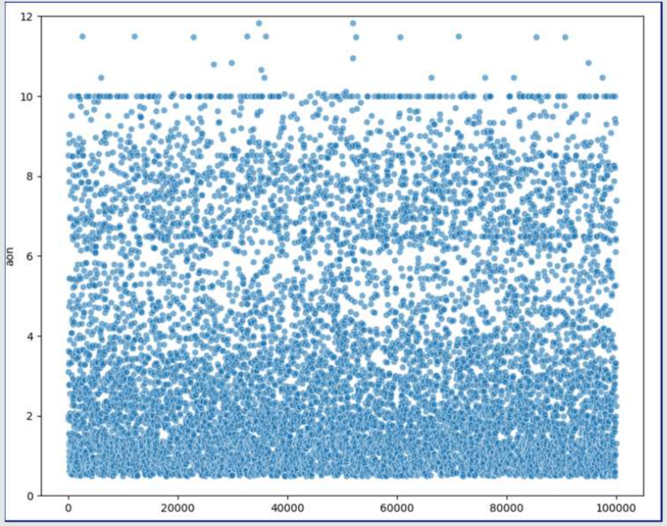
- Random Forests– The final Random forests train and test data accuracy is 95% and 92% respectively.



```
Train data result
              precision    recall  f1-score   support

           0       0.92      0.97      0.95        37
           1       0.97      0.92      0.95        38

    accuracy                           0.95        75
   macro avg       0.95      0.95      0.95        75
weighted avg       0.95      0.95      0.95        75

Test data result
              precision    recall  f1-score   support

           0       0.92      0.92      0.92        13
           1       0.92      0.92      0.92        12

    accuracy                           0.92        25
   macro avg       0.92      0.92      0.92        25
weighted avg       0.92      0.92      0.92        25
```

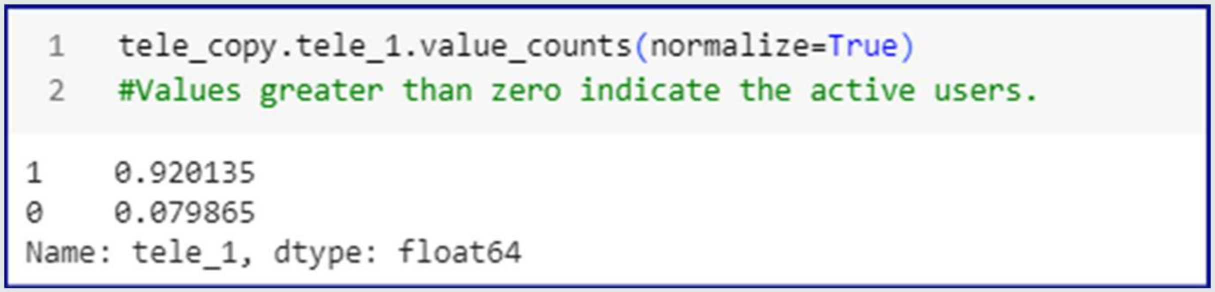
ANALYSIS

- Scatterplot – it allowed us to conclude that, majority of the customers were with the telecom company for less than 4 years. Therefore, customer retention beyond this period would be essential, as the longer a customer stays, the longer the customer tends to be responsive to new offers, billing plans, rate cards, etc., which prove as a source of long-term revenue for the telecom company.



ANALYSIS

- No. of Active users – The ratio of active users to the inactive ones is good at 92:08. This means, a large customer base has been actively utilizing the services of the telecom company. This ratio further needs to be compared to the industry standard in which the telecom company operates and in comparison, to other telecom companies in the same industry to get a more holistic view. From the Scatter Plot in the previous slide, it can be inferred that, this 92% active users consists mostly of customers who were with the telecom company for less than 4 years and therefore, steps must be taken to retain them beyond 4 years.



```
tele_copy.tele_1.value_counts(normalize=True)
#Values greater than zero indicate the active users.
```

```
1    0.920135
0    0.079865
Name: tele_1, dtype: float64
```