Telecom Churn –
Domain - Oriented Case Study

Problem Statement (Telco Domain)

In the telecom industry, customers are able to 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.**

In this project, you will analyze 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.

Understanding & Collection of Data

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).

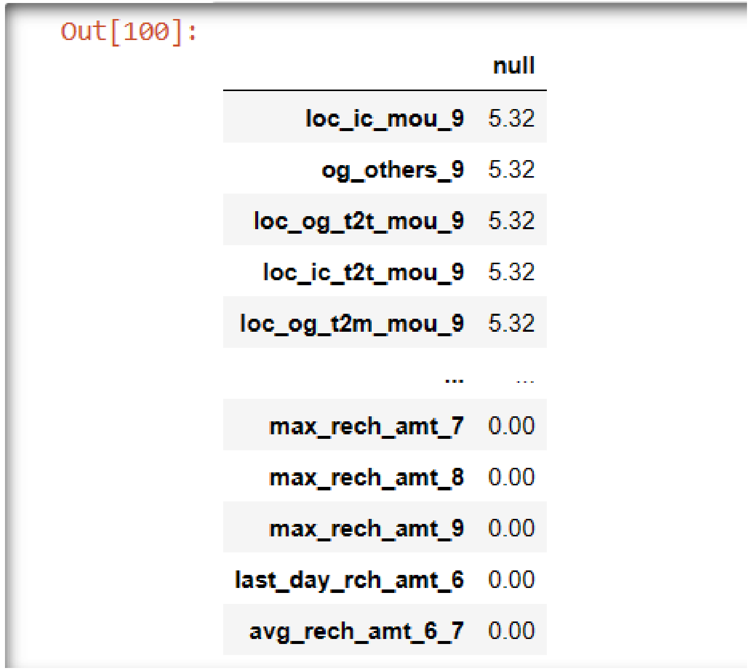
In the postpaid model, when customers want to switch to another operator, they usually inform the existing operator to terminate the services, and you directly know that this is an instance of churn.

However, in the prepaid model, 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 actually 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.

Data Preparation(2/3)

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.



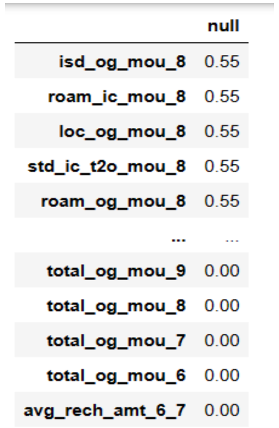
Out[100]:

	null
loc_ic_mou_9	5.32
og_others_9	5.32
loc_og_t2t_mou_9	5.32
loc_ic_t2t_mou_9	5.32
loc_og_t2m_mou_9	5.32
...	...
max_rech_amt_7	0.00
max_rech_amt_8	0.00
max_rech_amt_9	0.00
last_day_rch_amt_6	0.00
avg_rech_amt_6_7	0.00

Looks like MOU for all the types of calls for the month of September (9) have missing values together for any particular record.

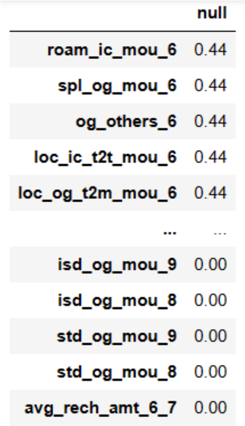
Lets check the records for the MOU for Sep(9), in which these coulmns have missing values together.

Data Preparation(2/2)



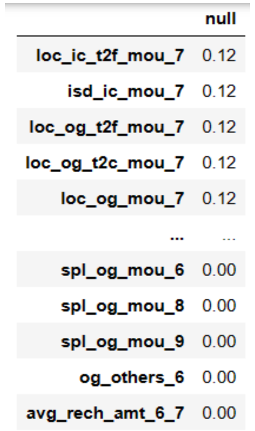
	null
isd_og_mou_8	0.55
roam_ic_mou_8	0.55
loc_og_mou_8	0.55
std_ic_t2o_mou_8	0.55
roam_og_mou_8	0.55
...	...
total_og_mou_9	0.00
total_og_mou_8	0.00
total_og_mou_7	0.00
total_og_mou_6	0.00
avg_rech_amt_6_7	0.00

Looks like MOU for all the types of calls for the month of Aug (8) have missing values together for any particular record. Lets check the records for the MOU for Aug(8), in which these coulmns have missing values together.



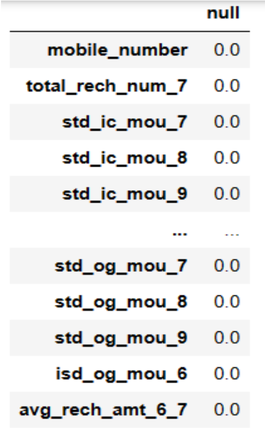
	null
roam_ic_mou_6	0.44
spl_og_mou_6	0.44
og_others_6	0.44
loc_ic_t2t_mou_6	0.44
loc_og_t2m_mou_6	0.44
...	...
isd_og_mou_9	0.00
isd_og_mou_8	0.00
std_og_mou_9	0.00
std_og_mou_8	0.00
avg_rech_amt_6_7	0.00

Looks like MOU for all the types of calls for the month of Jun (6) have missing values together for any particular record. Lets check the records for the MOU for Jun(6), in which these coulmns have missing values together.



	null
loc_ic_t2f_mou_7	0.12
isd_ic_mou_7	0.12
loc_og_t2f_mou_7	0.12
loc_og_t2c_mou_7	0.12
loc_og_mou_7	0.12
...	...
spl_og_mou_6	0.00
spl_og_mou_8	0.00
spl_og_mou_9	0.00
og_others_6	0.00
avg_rech_amt_6_7	0.00

Looks like MOU for all the types of calls for the month of July (7) have missing values together for any particular record. Lets check the records for the MOU for Jul(7), in which these coulmns have missing values together.



	null
mobile_number	0.0
total_rech_num_7	0.0
std_ic_mou_7	0.0
std_ic_mou_8	0.0
std_ic_mou_9	0.0
...	...
std_og_mou_7	0.0
std_og_mou_8	0.0
std_og_mou_9	0.0
isd_og_mou_6	0.0
avg_rech_amt_6_7	0.0

We can see there are no more missing values in any columns.
Checking percentage of rows we have lost while handling the missing values
round((1- (len(df.index)/30011)),2) – **0.07**
We can see that we have lost almost 7% records. But we have enough number of records to do our analysis.

Now 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.
Deleting all the attributes corresponding to the churn phase

Data Preparation(3/3)

In the filtered dataset except mobile_number and churn columns all the columns are numeric types. Hence, converting mobile_number and churn datatype to object.
Derive new features List the columns of total mou, rech_num and rech_amt.
decrease_mou_action :- This column indicates whether the minutes of usage of the customer has decreased in the action phase than the good phase.

decrease_rech_num_action :-This column indicates whether the number of recharge of the customer has decreased in the action phase than the good phase.

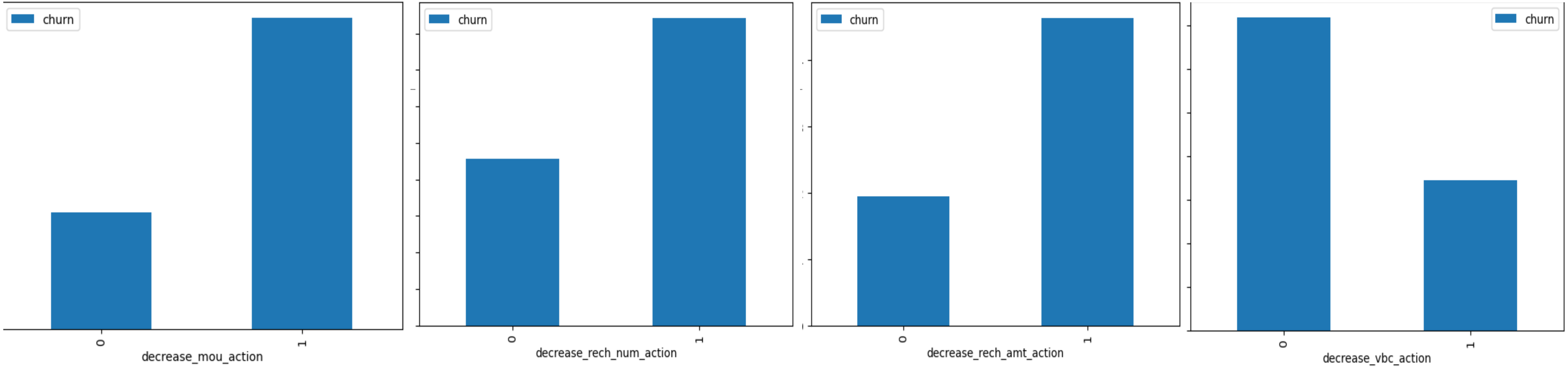
decrease_rech_amt_action:- This column indicates whether the amount of recharge of the customer has decreased in the action phase than the good phase.

decrease_arpu_action:- This column indicates whether the average revenue per customer has decreased in the action phase than the good phase.

decrease_vbc_action :- This column indicates whether the volume based cost of the customer has decreased in the action phase than the good phase.

Perform EDA - **Univariate analysis**

Churn rate on the basis whether the customer, decreased her/his MOU in action month, number of recharge in action month, amount of recharge in action month & volume based cost in action month



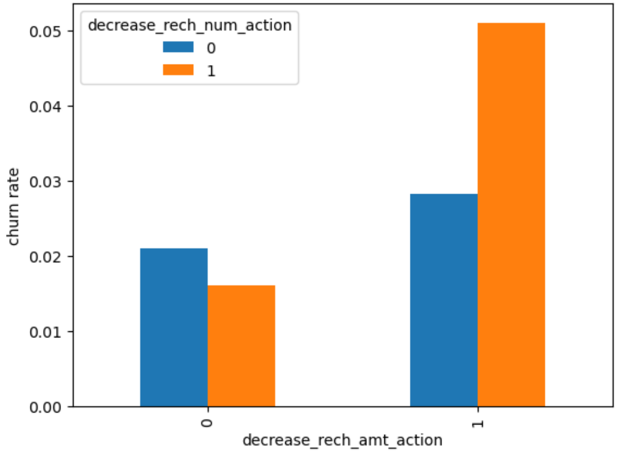
The churn rate is more for the customers, whose volume based cost in action month is increased. That means the customers do not do the monthly recharge more when they are in the action phase.

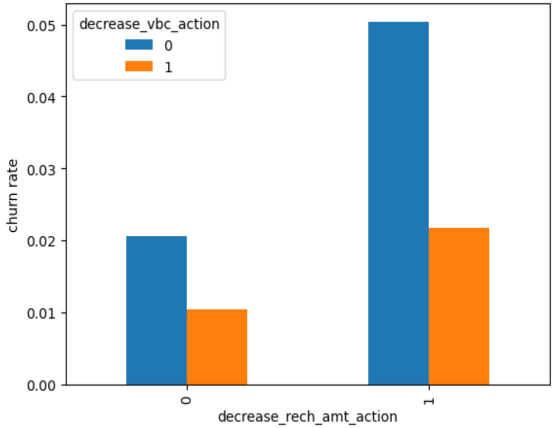
Average revenue per user (ARPU) for the churned customers is mostly densed on the 0 to 900. The higher ARPU customers are less likely to be churned.

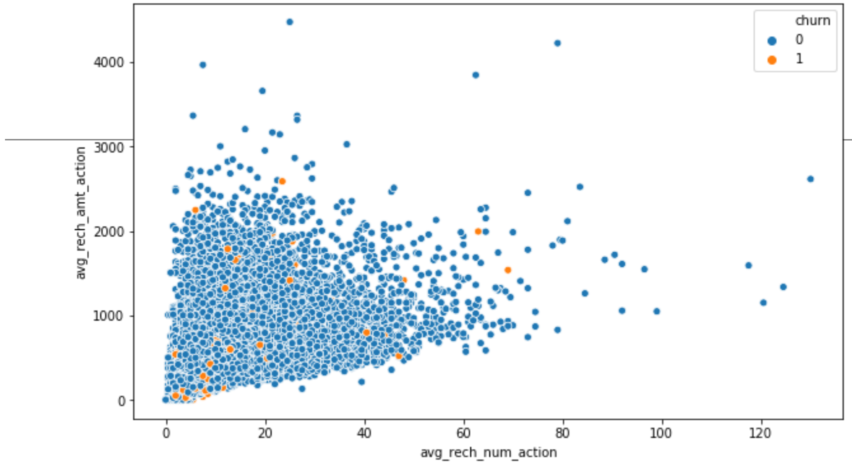
ARPU for the not churned customers is mostly densed on the 0 to 1000.

Perform EDA - **Bivariate analysis**

Analysis of churn rate by the decreasing recharge amount and number of recharge in the action phase, recharge amount and volume based cost in the action phase







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.

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.

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.

Dropping few derived columns, which are not required in further analysis

Feature Selection

Train-Test Split

Dealing with data imbalance using SMOTE(Synthetic Minority Oversampling Technique).

Feature Scaling

Build Models

Model with PCA

- Plotting scree plot. We can see that 60 components explain amost more than 90% variance of the data. So, we will perform PCA with 60 components.

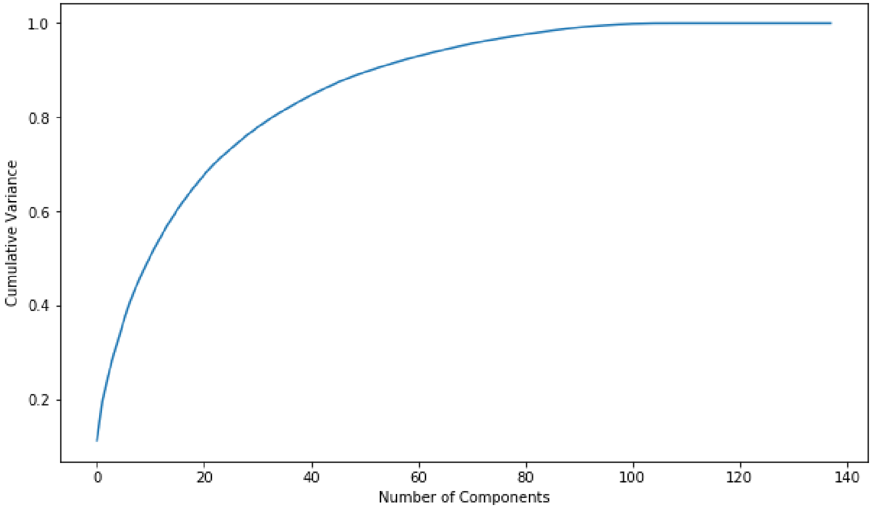
- We are only doing Transform in the test set not the Fit-Transform. Because the Fitting is already done on the train set. So, we just have to do the transformation with the already fitted data on the train set.

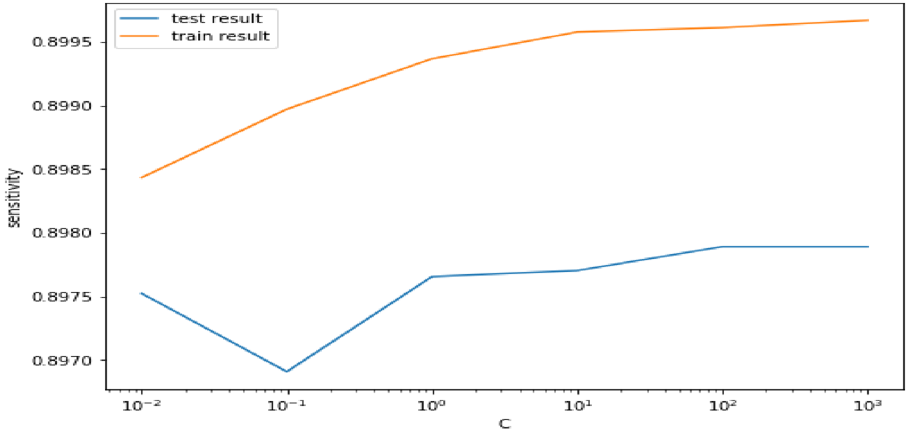
- We are more focused on higher Sensitivity/Recall score than the accuracy.

- Beacuse we need to care more about churn cases than the not churn cases. The main goal is to reatin the customers, who have the possiblity to churn. There should not be a problem, if we consider few not churn customers as churn customers and provide them some incentives for retaining them. Hence, the sensitivity score is more important here.

- Tuning hyperparameter C is the the inverse of regularization strength in Logistic Regression. Higher values of C correspond to less regularization.

- Plot of C versus train and validation scores. The highest test sensitivity is 0.8978916608693863 at C = 100





[[4452 896] [36 157]]

Validate & Measure Model Performance

Prediction on the train set

Confusion matrix –

```
[[17908  3517]
 [ 2154 19271]]
```

Accuracy:- 0.8676546091015169

Sensitivity:- 0.899463243873979

- **Prediction on the test set**
- Confusion matrix –

```
[[17908  3517]
 [ 2154 19271]]
```

Accuracy:- 0.8676546091015169

Sensitivity:- 0.899463243873979

Specificity:- 0.8358459743290548

Model summary

•Train set

- Accuracy = 0.86
- Sensitivity = 0.89
- Specificity = 0.83

•Test set

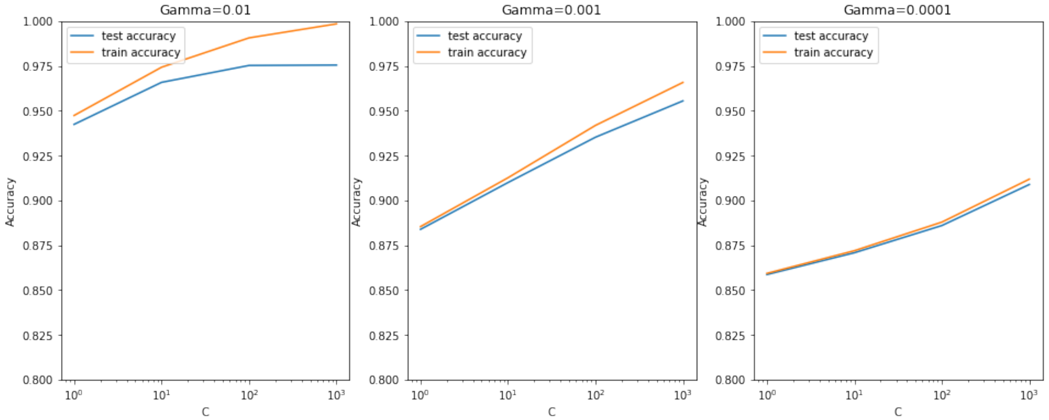
- Accuracy = 0.83
- Sensitivity = 0.81
- Specificity = 0.83

Overall, the model is performing well in the test set, what it had learnt from the train set.

Improve Model Performance

Support Vector Machine(SVM) with PCA

Plotting the accuracy with various C and gamma values



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.

Recomendations

1. 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).
2. Target the customers, whose outgoing others charge in July and incoming others on August are less.
3. 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.
4. Cutomers, whose monthly 3G recharge in August is more, are likely to be churned.
5. 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.
6. Cutomers decreasing monthly 2g usage for August are most probable to churn.
7. Customers having decreasing incoming minutes of usage for operators T to fixed lines of T for August are more likely to churn.
8. 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.