Telecom Churn – Domain Oriented Case Study

PROBLEM STATEMENT:

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 .To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.

BUSINESS OBJECTIVE:

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. To do this task well, understanding the typical customer behavior during churn will be helpful.

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.

Understanding and defining churn :

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).

Definitions of churn :

There are various ways to define churn, such as:

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 customer behavior during churn :

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.

Data dictionary :

	Acronyms	Descriptions
2	MOBILE_NUMBER	Customer phone number
3	CIRCLE_ID	Telecom circle area to which the customer belongs to
4	LOC	Local calls - within same telecom circle
5	STD	STD calls - outside the calling circle
6	IC	Incoming calls
7	OG	Outgoing calls
8	T2T	Operator T to T, i.e. within same operator (mobile to mobile)
9	T2M	Operator T to other operator mobile
10	T2O	Operator T to other operator fixed line
11	T2F	Operator T to fixed lines of T
12	T2C	Operator T to it's own call center
13	ARPU	Average revenue per user
14	MOU	Minutes of usage - voice calls
15	AON	Age on network - number of days the customer is using the operator T network
16	ONNET	All kind of calls within the same operator network
17	OFFNET	All kind of calls outside the operator T network
18	ROAM	Indicates that customer is in roaming zone during the call
19	SPL	Special calls
20	ISD	ISD calls
21	RECH	Recharge
22	NUM	Number
23	AMT	Amount in local currency
24	MAX	Maximum
25	DATA	Mobile internet
26	3G	3G network
27	AV	Average
28	VOL	Mobile internet usage volume (in MB)
29	2G	2G network
30	PCK	Prepaid service schemes called - PACKS
31	NIGHT	Scheme to use during specific night hours only
32	MONTHLY	Service schemes with validity equivalent to a month
33	SACHET	Service schemes with validity smaller than a month
34	*.6	KPI for the month of June
35	*.7	KPI for the month of July
36	*.8	KPI for the month of August
37	*.9	KPI for the month of September
38	FB_USER	Service scheme to avail services of Facebook and similar social networking sites
39	VBC	Volume based cost - when no specific scheme is not purchased and paid as per usage

The data dictionary contains meanings of abbreviations. Some frequent ones are loc (local), IC (incoming), OG (outgoing), T2T (telecom operator to telecom operator), T2O (telecom operator to another operator), RECH (recharge) etc.

The attributes containing 6, 7, 8, 9 as suffixes imply that those correspond to the months 6, 7, 8, 9 respectively.

Data preparation :

1. Filter high-value customers

As mentioned above, you need to predict churn only for high-value customers. Define high-value customers as follows: Those who have recharged with an amount more than or equal to X, where X is the 70th percentile of the average recharge amount in the first two months (the good phase).

After filtering the high-value customers, you should get about 30k rows.

2. Tag churners and remove attributes of the churn phase

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. The attributes you need to use to tag churners are:

total_ic_mou_9

total_og_mou_9

vol_2g_mb_9

vol_3g_mb_9

After tagging churners, remove all the attributes corresponding to the churn phase (all attributes having ' _9', etc. in their names).

DATA CLEANING:

1 Check and handle duplicate data.

2 Check and handle NA values and missing values.

3 Drop columns, if it contains large amount of missing values and not useful for the analysis.

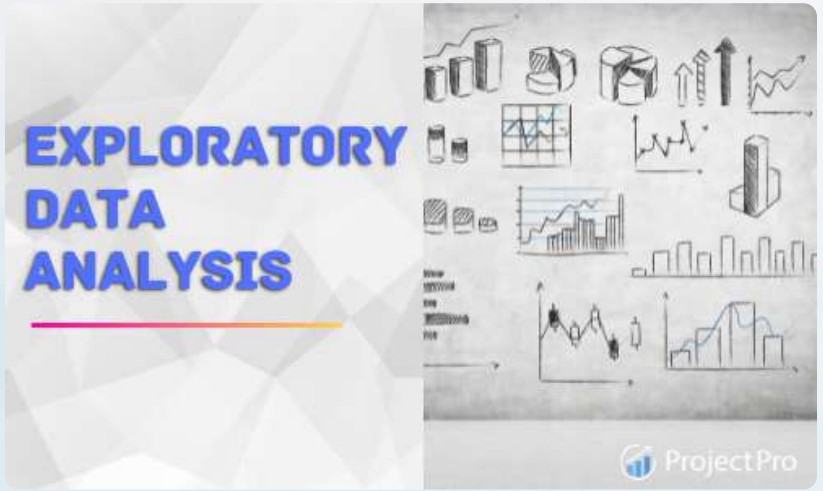
4 Imputation of the values, if necessary.

5 Check and handle outliers in data.

Exploratory Data Analysis:

Univariate data analysis :

- value count
- distribution of variable



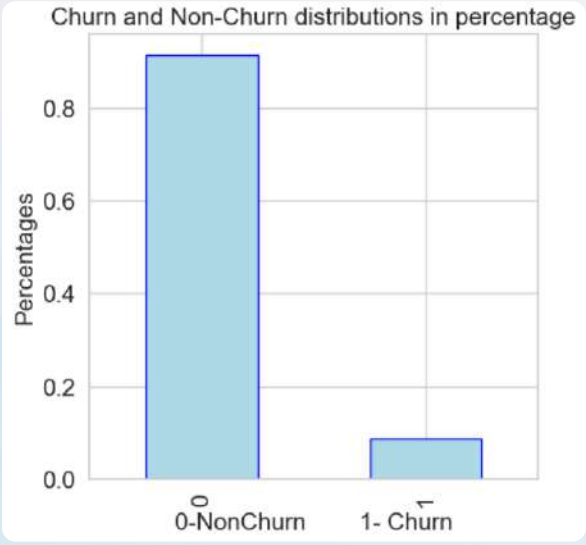
Bivariate data analysis:

- correlation coefficients
- pattern between the variables

• Very high conversion rates for lead sources ‘Reference’ and ‘Welingak Website’.

• Most leads are generated through ‘Direct Traffic’ and ‘Google’

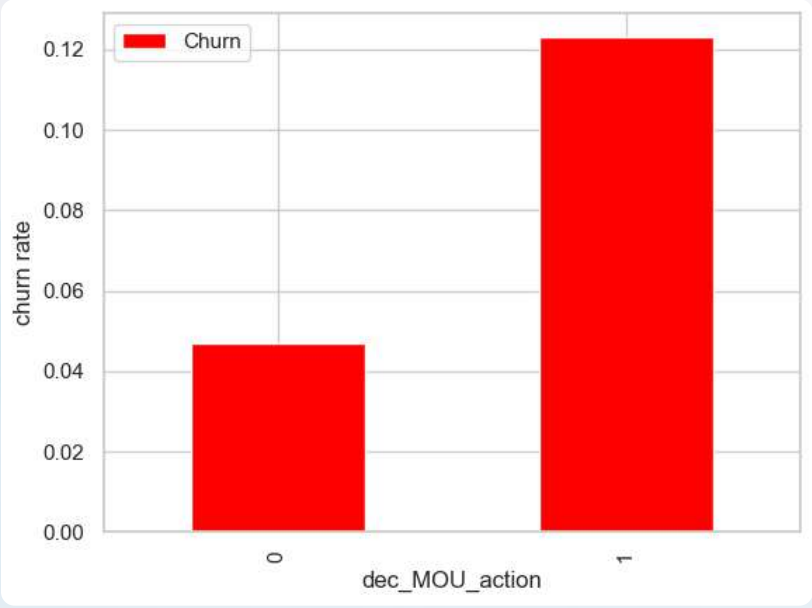
Churn and Non-Churn distributions in percentage:



This tells us that 8.6% customers have churned. Which indicates class imbalance, we will take care of it at later point by using SMOTE.

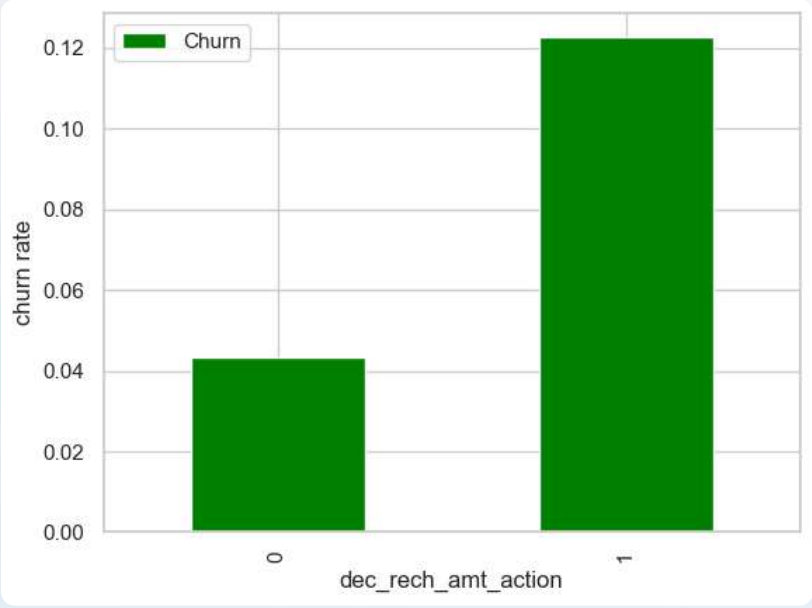
Churn rate based on whether the customer decreased her/his MOU in the action month:

We can see that the churn rate is more for the customers, whose minutes of usage(mou) decreased in the action phase than the good phase.



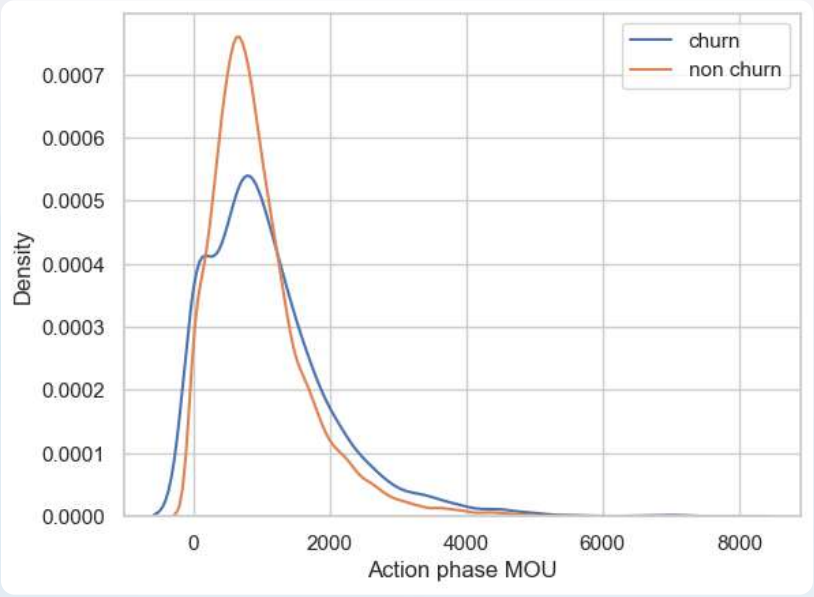
Churn rate on the basis of whether the customer decreased the amount of recharge in the action month:

Here also we see the same behavior. The churn rate is more for the customers, whose amount of recharge in the action phase is lesser than the amount in the good phase.



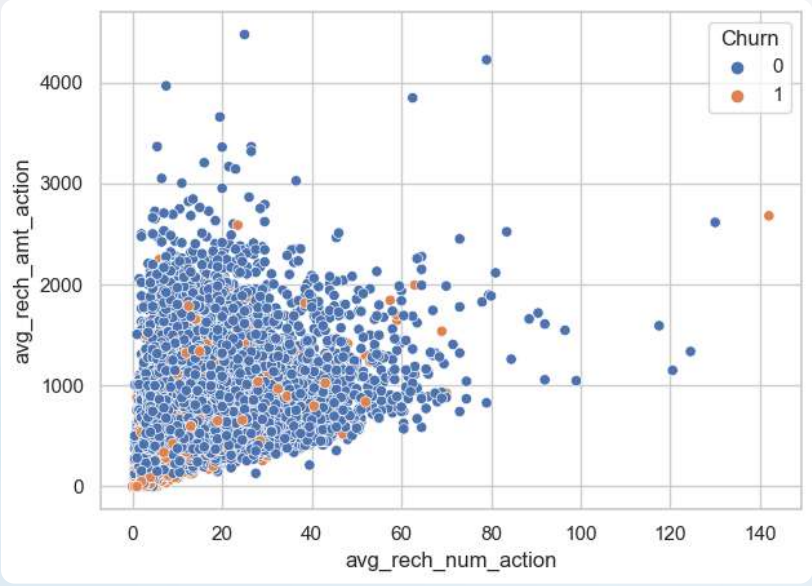
Analysis of the minutes of usage MOU (churn and not churn) in the action phase:

Customers that churn tend to have minutes of usage (MOU) that range from 0 to 2500. Higher the MOU, the lesser the churn probability



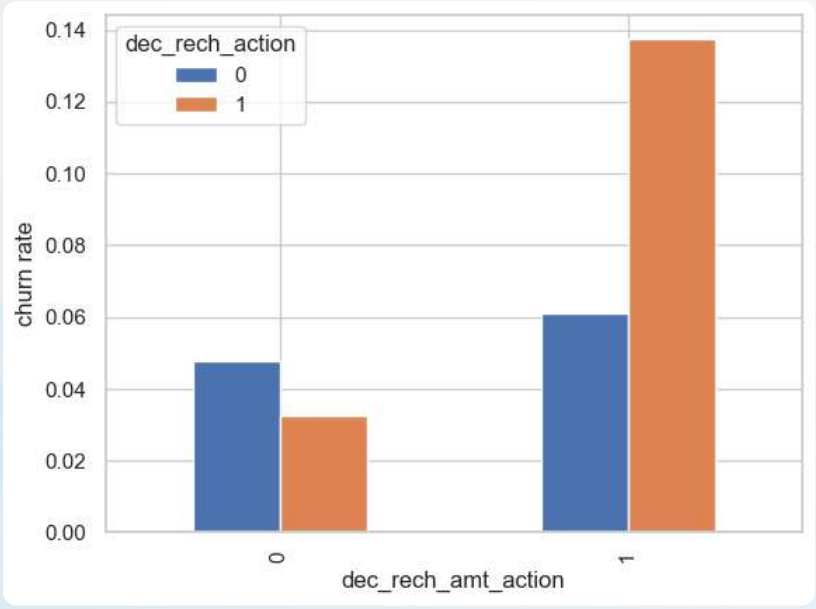
Analyzing recharge amount and number of recharge in action month:

We can see from the above pattern that the recharge number and the recharge amount are almost proportional. Higher the number of recharge, Higher is the amount of the recharge.



Analyzing churn rate WRT the decreasing recharge amount and number of recharge during the action phase:

We can see from the above plot, that the churn rate is higher for the customers, whose recharge amount as well as the number of recharges have decreased in the action phase when compared to the good phase.



Train- Test split:

Train Test Split

In [73]:
```
from sklearn.model_selection import train_test_split

# Putting feature variables into X
X = new_DF.drop(['Churn'], axis=1)

# Putting target variable to y
y = new_DF.pop('Churn')

# Splitting data into train and test set 70:30
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7, test_size=0.3, random_state=100, stratify=y)
```

In [74]:
```
X_train.shape, X_test.shape, y_train.shape, y_test.shape
```

Out[74]:
```
((20790, 137), (8910, 137), (20790,), (8910,))
```

❖ We have split train data to 70% and test data is 30%.

COMPLETE MODEL STATS:

	Model	Recall	Test Accuracy	Roc_auc_score
1	Decision Tree with PCA	0.89	0.83	0.77
0	Logistic Regression with PCA	0.87	0.83	0.88
3	Logistic without PCA	0.82	0.79	0.76
2	Random Forest with PCA	0.70	0.87	0.88

Conclusion and STRATEGY ahead:

1. From EDA, we observed that there is a considerable drop in recharge, call usage and data usage in the 8th month which is the Action Phase. Below are the important features: loc_og_t2m_mou_7, total_og_mou_6, loc_og_t2t_mou_7, roam_ic_mou_7, onnet_mou_7, arpu_7,loc_og_t2c_mou_7, onnet_mou_8, roam_og_mou_8, arpu_6
2. Average revenue per user in the 7th month plays a vital role in deciding churn. A sudden drop in it might indicate that the customer might be thinking about churning and appropriate actions should be taken.
3. Local Minutes of usage (outgoing) are the most affecting features on the customer churn.
4. Roaming Minutes of usage (incoming & outgoing) are also affecting features on the customer churn.
5. Total minutes of usage for outgoing is also an important factor affecting the churn.

Following strategies can be incorporated:

• A sudden drop in Local Minutes of usage might be because of unsatisfactory customer service because of poor network or unsuitable customer schemes/plans. Efforts shall be made to provide a better network and focus on customer satisfaction.

• Based on the usage / last recharge/ net usage, routine feedback calls should be made for customer satisfaction and services that can understand their grievances & expectations. Appropriate action should be taken to avoid them from churning.

• Various attractive offers can be introduced to customers showing a sudden drop in the total amount spent on calls & data recharge in the action phase to lure them.

• Customized plans should be provided to such customers to stop them from churning.

• Promotional offers can also be very helpful.