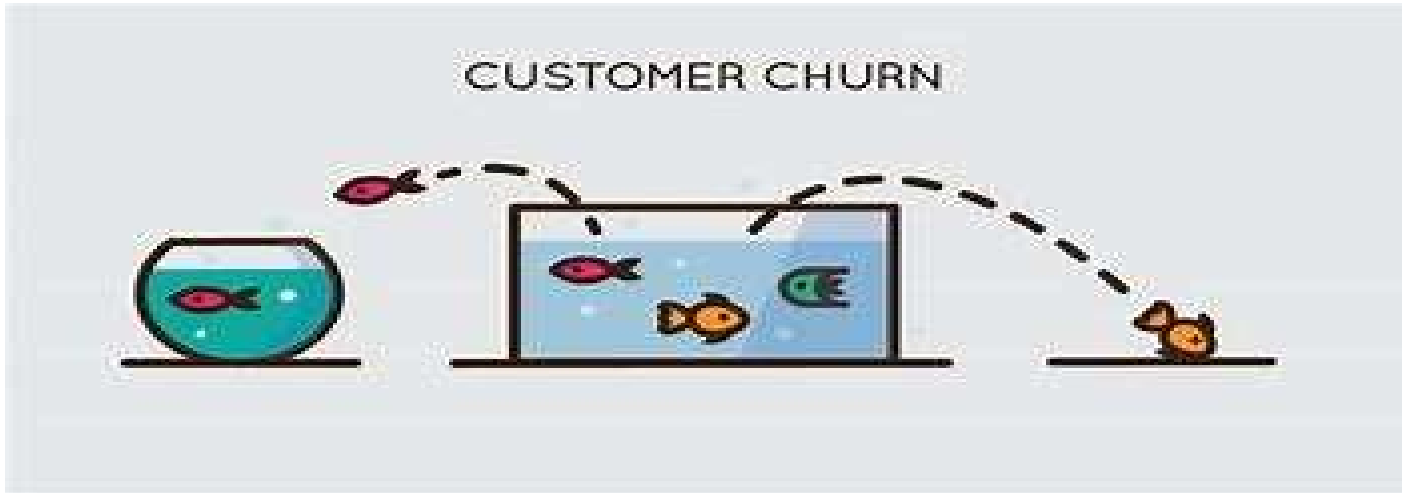# EDA ON TELECOM CHURN

(Pratyush Kumar Rath)

Introduction:

CUSTOMER CHURN

Customer churn means shifting from one service provider to its competitor in the market. Customer churn is one of the biggest fears of any industry, particularly for the telecom industry. With an increase in the number of telecom service providers in South Asia, the level of competition is quite high.

Losing customers is costly for any business. As per the authors of "Leading on the Edge of Chaos". A 2% decrease in customer churn is equivalent to 10% reduction in costs. Moreover, as per the White House Office of Consumer Affairs, acquiring new customers is 6–7 times more expensive than retaining an old one. Identifying unhappy customers early, taking into consideration their values and the risk to churn, give you a chance to offer them incentives to stay. Examples of customer churn include cancelation of a subscription, closure of an account, non-renewal of a contract or service agreement, or use another service provider.

Although there are many reasons for customer churn, some of the major reasons are service dissatisfaction, costly subscription, and better alternatives. The telecom service providers strive very hard to sustain in this competition. So to sustain this competition they often try to retain their customers than acquiring new ones as it proved to be much costlier. Hence predicting churn in the telecom industry is very important. To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.

Here are steps of data science life cycle to make sure wise data-driven decisions have been made to fight against the customer churn:
1- Understand the business model, requirements, and risks.
2- Collect available customer behaviour, transactions, demographics data, and usage patterns.

3- Formulate initial hypotheses based on domain knowledge and exploratory analysis that will assist in defining metrics to tackle these hypotheses.

4- Do the required data processing to have a cleaned dataset.

5- Assess the results to ensure that the model satisfies the original business goals.

6- Derive conclusion by the best processing of the data.

Business Understanding:

Organizations spend huge amounts of money to acquire customers. Even before a customer is acquired, you must gain brand and product awareness and customer consideration. This requires lots of effort throughout the sales funnel, using lots of marketing and sales resources. Once a customer is acquired, they should be held on to dearly (at least the ones with high lifetime value). Knowing which customers have a high propensity to churn is critical to targeted retention efforts

Life Cycle Of Customer Churn Prediction Project:
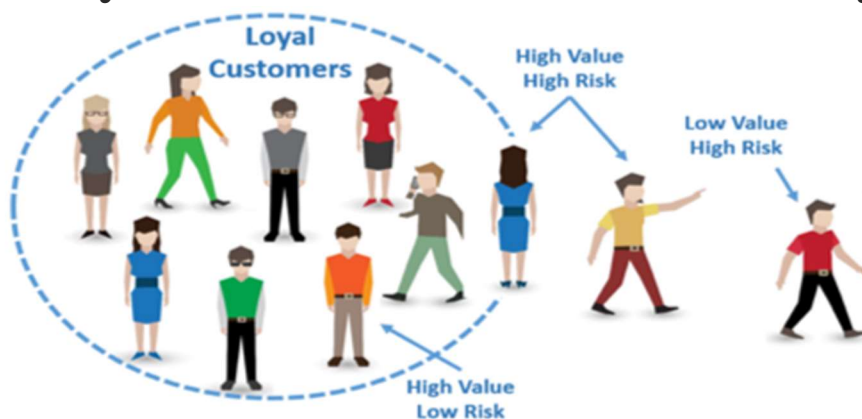
Data Acquisition:

In our problem, we use a dataset file with .csv extension that is publicly available on GitHub. However, to optimize our model, more data might be required through other channels such as web scraping, API, Open Data, or Databases.

Introduction To The Dataset:

Most Mobile operators have historical records on which customers ended up churning and which continued using their services.. Here is a sample of the first 5 observations:

State	Account Length	Area Code	Phone	Int'l Plan	VMail Plan	VMail Message	Day Mins	Day Calls	Day Charge	Eve Mins	Eve Calls	Eve Charge	Night Mins	Night Calls	Night Charge	Intl Mins	Intl Calls	Intl Charge	CustServ Calls	Churn?
KS	128	415	382-4657	no	yes	25	265.1	110	45.07	197.4	99	16.78	244.7	91	11.01	10	3	2.7	1	False.
OH	107	415	371-7191	no	yes	26	161.6	123	27.47	195.5	103	16.62	254.4	103	11.45	13.7	3	3.7	1	False.
NJ	137	415	358-1921	no	no	0	243.4	114	41.38	121.2	110	10.3	162.6	104	7.32	12.2	5	3.29	0	False.
OH	84	408	375-9999	yes	no	0	299.4	71	50.9	61.9	88	5.26	196.9	89	8.86	6.6	7	1.78	2	False.
OK	75	415	330-6626	yes	no	0	166.7	113	28.34	148.3	122	12.61	186.9	121	8.41	10.1	3	2.73	3	False.

It's a relatively small dataset, only 3,333 records, where each record uses the first 20 attributes to describe the profile of a certain customer and the last attribute to label this customer. This dataset of an unknown US mobile operator. Here are the data dictionary:

1- State: the US state in which the customer resides, indicated by a two-letter abbreviation

2- Account Length: the number of days that this account has been active

3- Area Code: the three-digit area code of the corresponding customer's phone number

4- Phone: the remaining seven-digit phone number

5- Int'l Plan: whether the customer has an international calling plan: yes/no

6- VMail Plan: whether the customer has a voice mail feature: yes/no

7- VMail Message: presumably the average number of voice mail messages per month

8- Day Mins: the total number of calling minutes used during the day

9- Day Calls: the total number of calls placed during the day

10- Day Charge: the billed cost of daytime calls

11- Eve Mins: the total number of calling minutes used during the evening

12- Eve Calls: the total number of calls placed during the evening

13- Eve Charge: the billed cost of evening time calls

14- Night Mins: the total number of calling minutes used during the night

15- Night Calls: the total number of calls placed during the night

16- Night Charge: the billed cost of night time calls

17- Intl Mins: the total number of international minutes

18- Intl Calls: the total number of international calls

19- Intl Charge: the billed cost for international calls

20- Cust Service Calls: the number of calls placed to Customer Service

21- Churn: whether the customer left the service: true/false

Data Preparation:

These four criteria can be used to ensure the quality of our dataset:

- Complete: There are no missing values or nulls in our data. However, some dimensions are missing specially the demographics. Characteristics such as race, ethnicity, gender, age, education, profession, occupation, income level, and marital status can increase the accuracy of the prediction.

- Clean: Phone numbers are unique values so that it can be anonymously encoded and used as row index. Also, the nominal or logical attributes such as plans and our target variable (Churn? :) need to be converted into numerical values, which are a required format to be used as an input to the machine learning algorithms.

- Accurate: Outliers and values that don't make sense need to be visualized and discussed with the stakeholders whether to include these data or not.

- Engineered Features: Some features need to be added to tackle our hypothesis. Although it might not help as predictive power, but it could assist in data visualization by dimensionality reduction. The first feature, can be called Service Call Rate, which is defined as Cust Service Calls/ Account Length. The second one, Total Charge = Day Charge + Eve Charge + Night Charge + Intl Charge.

Data Visualisation:

- Graphical Representation Of The Results: This step involves presenting the dataset to the target audience in the form of graphs summary tables, maps,

and diagrams. This is also an essential step as the result analyzed from the dataset should be interpretable by the business stakeholders, which is one of the major goals Of EDA. Most Of the graphical analysis techniques include Line chart, Bar chart, Scatter plot, Area plot, and stacked plot Pie chart, Table chart, Polar chart, Histogram, Lollipop chart etc.

- Correlation Among Variables: In words, the statistical technique that examines the relationship and explains whether, and how strongly, pairs of variables are related to one another is known as correlation. Correlation answers questions such as how one variable changes with respect to another. If it does change, then to what degree or strength? Additionally, if the relation between those variables is strong enough, then we can make predictions for future behaviour.

Conclusion:

From the given data and after performing EDA and comparison with the all the elements we say that there are some factors which company should take care in consideration. If the churn rate of company is 50% then the company will be going to shut in two years. As company with churn rate 25% then it will shut in four years so Churn is the major factor to be taken in consideration. By analysing the data, we found some major conclusions which may considered with the churning rate of the company. These results include high pricing of international calls, low customer support system, quality of voice mail and also the network issue. It can be beneficial for the company if it looks into the considerations and resolve issues and improve itself.

THANK YOU