**Customer\_churn\_large\_dataset**

This dataset contains information related to customers and their interactions with a service provider.

* CustomerID: A unique identifier for each customer.
* Name: The name or identifier of each customer.
* Age: The age of each customer.
* Gender: The gender of the customer, with two categories, "Male" and "Female."
* Location: The location or city where the customer is located. It has categorical values like "Los Angeles," "New York," and "Miami."
* Subscription\_Length\_Months: The length of the customer's subscription in months. This can indicate how long a customer has been using the service.
* Monthly\_Bill: The amount billed to the customer on a monthly basis.
* Total\_Usage\_GB: The total amount of data or usage in gigabytes, which the customer has consumed.
* Churn: A binary variable indicating whether the customer has churned (1 for churned, 0 for not churned). Churn refers to customers who have discontinued their subscription or stopped using the service

In this dataset the dependent variable(y) is “Churn” whole the others are independent variables.

**Data visualization**

* I have used box plot to visualize the outliers in Monthly bill, total usage and subscription length months.
* A bar chart to visualize the average total usage in gigabytes for different gender categories within various locations. It provides insights into the distribution of data usage across gender and location categories.
* A set of histograms are used to visualize the distribution of the specified features, including "Age," "Location," "Subscription\_Length\_Months," "Monthly\_Bill," and "Total\_Usage\_GB." The histograms provide insights into the distribution and characteristics of these features, helping you understand the data better.
* A bar plot has been plotted to visualize the relationship between customer "Age" and the target variable "Churn" while distinguishing between "Gender" categories. This helps you understand how age and gender relate to customer churn, providing insights for further analysis and decision-making.

**Data Pre-processing/ Feature Engineering**

* We have checked for null values the dataset doesn’t seem to be having any null values.
* We have also checked whether the data has duplicate values.
* Age Binning: Binsare created from the 'Age' column to capture age-related patterns.Bins are :'Young,' 'Adult,' 'Middle-aged,' and 'Senior.' Then I have label encoded these bins.
* Other categorical columns like ‘Gender’ and ‘Location’ have also been label encoded using Label Encoder
* Subscription Length in Years: Convert 'Subscription\_Length\_Months' into 'Subscription\_Length\_Years' to get a more interpretable feature.
* Total Monthly Cost:Calculate the total monthly cost by multiplying 'Monthly\_Bill' by 'Subscription\_Length\_Months.'
* Monthly Usage per Year: Calculate 'Monthly\_Usage\_Per\_Year' by dividing 'Total\_Usage\_GB' by 'Subscription\_Length\_Years.' This feature could provide insights into usage patterns.
* I have used Min-max scaler for scaling features to a specific range, , [0, 1]

**Split the Data**

* Divide the dataset into training and testing sets to train the model on one part and evaluate its performance on another.
* I have divided data as 80% for training the model and 20% for testing the model using train\_test\_split.

**Model Building/Optimization**

* I have used various models to compare their accuracy based on precision,f1 score,recall etc.
* The models used are Logistic Regression as it works well for a classification problem especially Binary classification, XG Boost and RandomForest Classifier.
* I have also applied Cross validation and GridSearch Cv for tuning my model and optimizing it

**Deployment**

* I have used Flask to deploy the model.