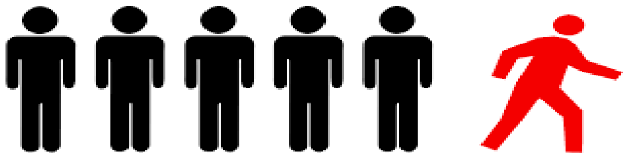
CUSTOMER CHURN PREDICITION



Telecom Industry

manisha_md@yahoo.com

Background :

- Telecom companies face major challenge with customer churn, as customers switch to alternate provider due to various reasons like lower cost, multi (combo) service offerings, marketing promotions by competitors, etc.
- Identifying these potential customers early on who may voluntarily churn and providing them retention incentives in form of discounts & combo offers will help the organization to retain those customers and reduce revenue loss.
- The company can also internally study any possible operational causes and improve its product offerings.
- Proactive actions will prevent the loss of revenue for the company and will improve / retain the market share among the industry peers in terms of the number of active subscribers.

Objective:

- The objective is to predict to a high accuracy, in advance the customers who may attrite from the existing service provider in near future.
- Analyze data using Exploratory Data Analysis and building model using Logistics Regression , XGBoost .
- Recommend product strategies to business team based on analysis of product offerings that will help in retaining the customer based on available data.

Dataset Description :

- Data consists of 7043 fictional customers who belong to various demographics (single; with dependents; senior citizen) subscribe to different products offerings (internet service; phone line; streaming TV; streaming movies; online security) from a telecom company located in one of the US states.
- 33 Independent variables with customer account information (contract; payment methods; location; charges)
- Dependent Target variable: “Churn”
- Churn Rate (Baseline) is 26.5%
- Dataset source: https://www.kaggle.com/datasets/becksddf/churn-in-telecoms-dataset

CustomerID
Count
Country
State
City
Zip Code
Lat Long
Latitude
Longitude
Gender
Senior Citizen
Partner
Dependents
Tenure Months
Phone Service
Multiple Lines
Internet Service
Online Security
Online Backup
Device Protection
Tech Support
Streaming TV
Streaming Movies
Contract
Paperless Billing
Payment Method
Monthly Charges
Total Charges
Churn Value
Churn Score
CLTV
Churn Reason
Churn Label

Enriching the dataset: Data Cleaning

1. Checking the data types of all the columns
2. Check the descriptive statistics of numeric variables
3. Create a copy of base data for manipulation & processing
4. Removing columns not required for processing - ChurnValue, ChurnScore, CLTV, ChurnReason, Country, ChurnLabel, CustomerID, Count, State, LatLong, Latitude, Longitude
5. Convert all the categorical variables into dummy variables for data processing

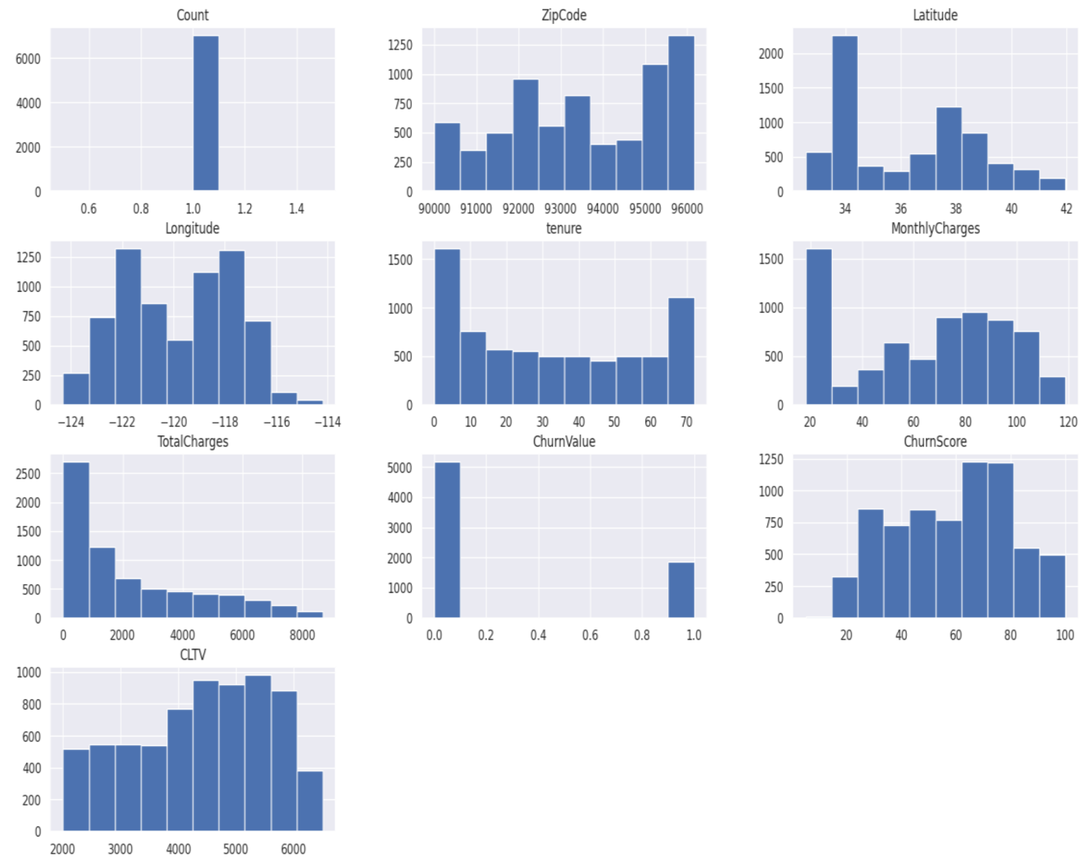
	ZipCode	tenure	MonthlyCharges	TotalCharges	City_Acampo	City_Acton	City_Adelanto	City_Adin	City_Agoura Hills	City_Aguanga	...
0	90003	2	53.85	108.15	0	0	0	0	0	0	...
1	90005	2	70.70	151.65	0	0	0	0	0	0	...
2	90006	8	99.65	820.50	0	0	0	0	0	0	...
3	90010	28	104.80	3046.05	0	0	0	0	0	0	...
4	90015	49	103.70	5036.30	0	0	0	0	0	0	...

5 rows × 1176 columns

Exploratory Data Analysis:

Study of Distribution :

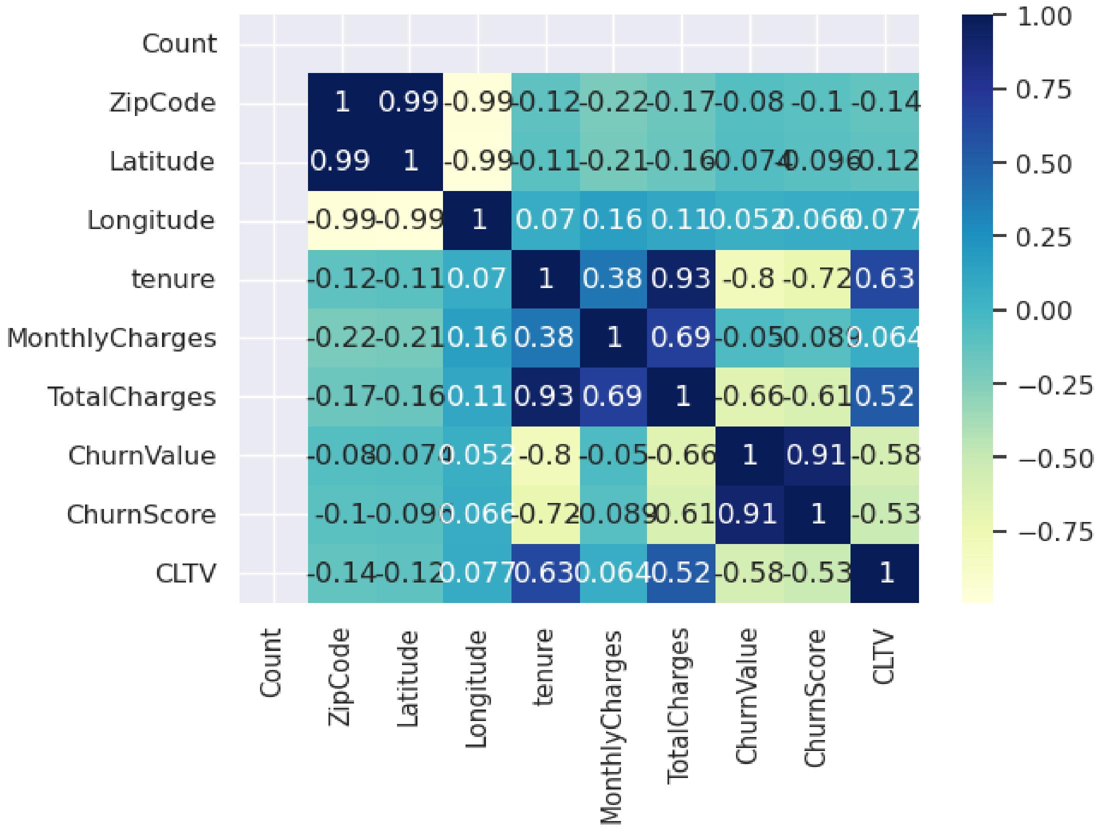
Helps in understanding the dataset better, whlie scaling



Correlation Heat Map:

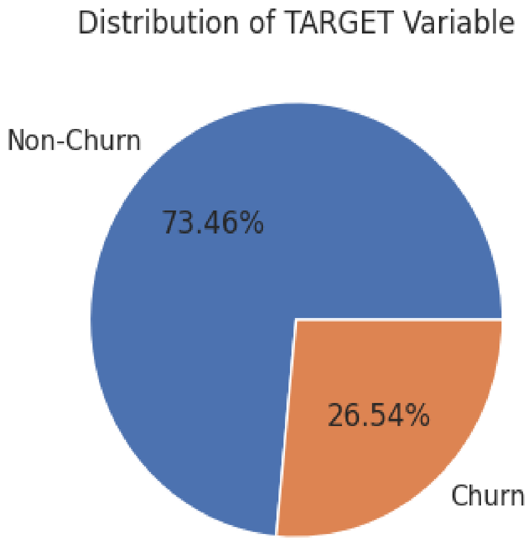
Variables close to 1 is highly correlated

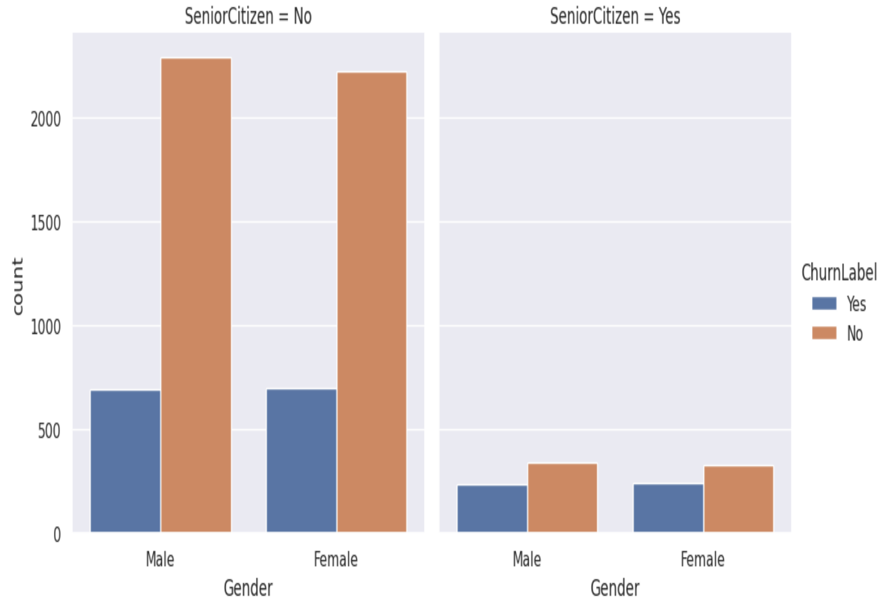
EX: Totalcharges, Tenure, ChurnValue, ChurnScore

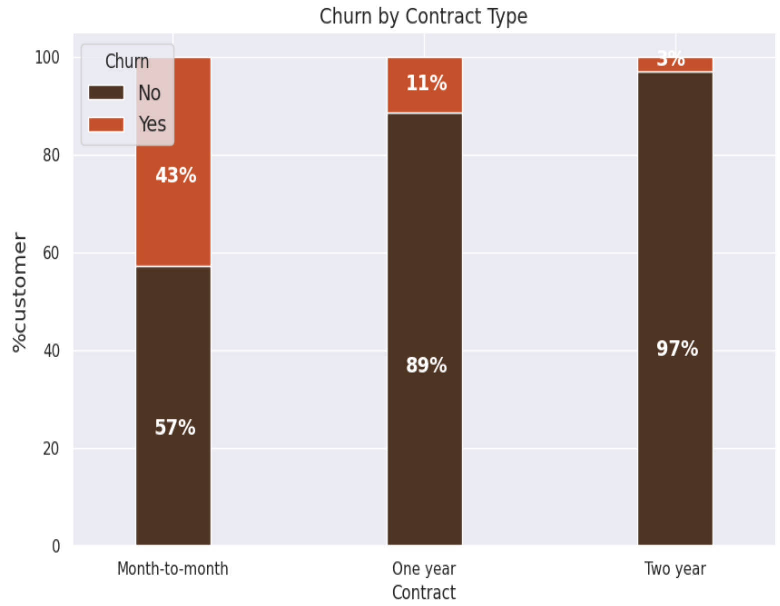


Churn Distributions:

- Data distribution of Target variable results as imbalanced data, ratio = 73.46%.
- Senior Citizens customers are less likely to Churn.
- Contract type with Long tenure are less likely to Churn.
- “Tenure” and “Contract” Type (monthly; 1 year; 2 year) are the most important variables

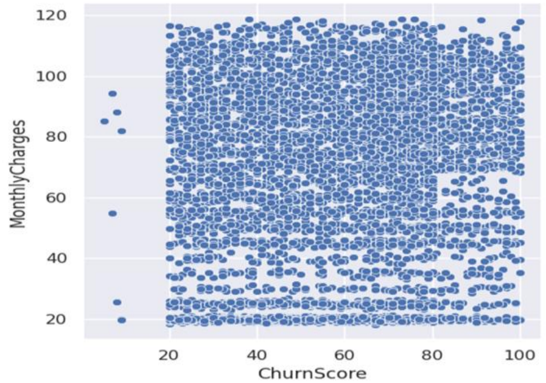




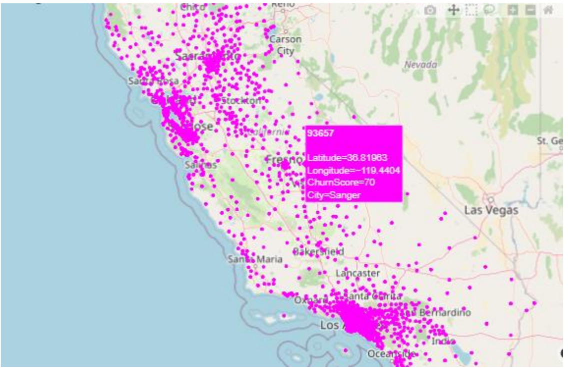


Data visualisation:

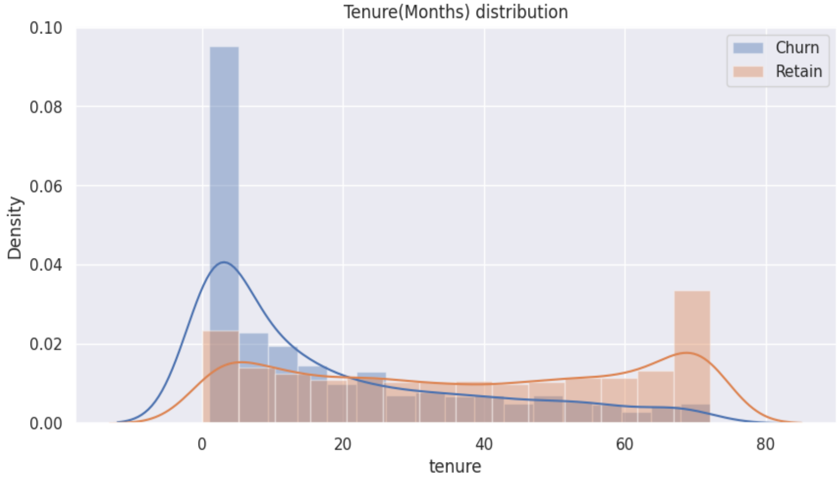
- Scatter plot between “ChurnScore” and “MonthlyCharges” showing Dense distribution of ChurnScore above 60 monthly charges.



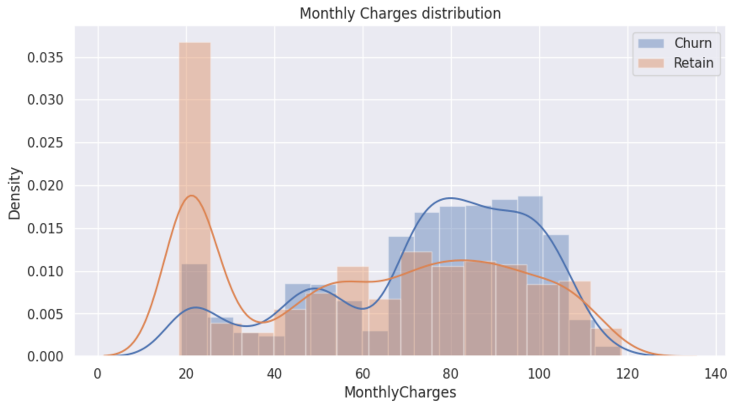
- Scatter map plot to shows High ChurnScore customers location and City

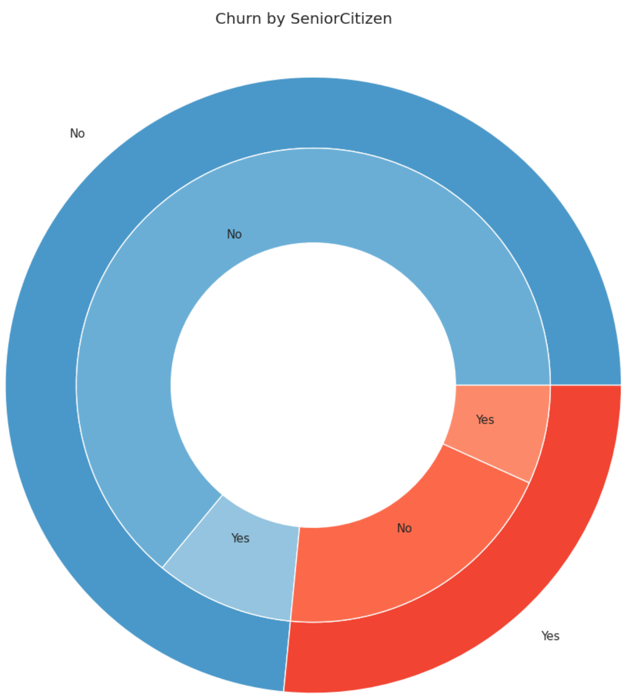


- As tenure increases customer retention is more, churn is high for short tenure.



- Similarly Monthly Charges between 60 to 120 retention increases then decrease charges are less customer retention is high.



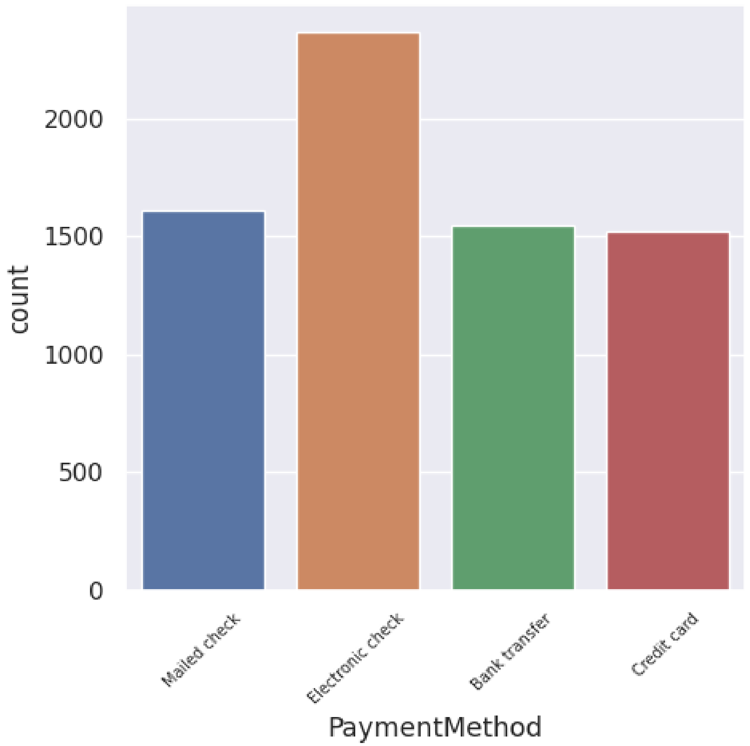


Churn by Senior Citizen:

Senior citizens are less likely to churn

Payment Method: Electronic Check shows

High churn ratio.



Conclusions :

- ✓ Non senior Citizens are high churners.
- ✓ Customer having Short tenure are high churners.
- ✓ Customer with Low Monthly charges are less churners.
- ✓ Electronic check medium are the highest churners.
- ✓ No Online security, No Tech Support category are high churners.
- ✓ Imbalanced dataset and concluded as Classification problem (as churn column is shown)

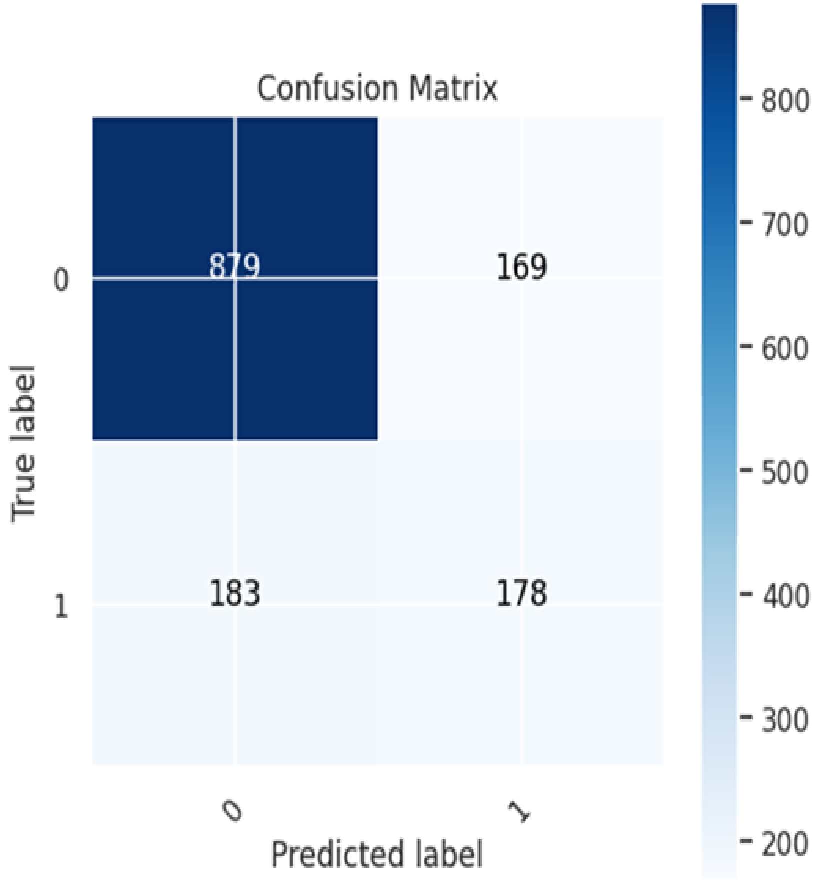
Data Modelling :

- ✓ Imbalanced dataset and conclude as Classification problem (as churn column is shown)
- ✓ For this Classification problem, ML Algorithms used
  1) XGBoost and
  2) Logistic Regression
- ✓ Confusion Matrix.
- ✓ **Result of two types:**

1) <u>Not handled class imbalance problem :</u>
   - ✓ Segregation of Categorical and numerical features
   - ✓ Split dataset into Train and test Data , standardization
   - ✓

Models	**Accurarcy**
XGBoost	**65.80%**
Logistic Regression	**75.01%**

Conclusion: Logistic Regression result accuracy is better than XGBoost.



Conclusion:

2) Class imbalance problem Handled
   - ✓ Applying SMOTE technique to handle class imbalance problem.
   - ✓ Split dataset into Train and test Data , standardization.
   - ✓ Applying PCA - Dimensionality Reduction on training and test data.
   - ✓ Logistic Regression result accuracy is 75.01%.

- ❑ **Conclusion: With PCA - Dimensionality Reduction , we couldn't see any better results, hence finalizing the model by Logistic Regression.**