**CUSTOMER CHURN ANALYSIS**

## Abhijit Sarkar

**Let’s Know What is Customer Churn?**

##### **The Customer Churn is also known as Customer Attrition**.

Customer churn is when a company’s customers stop doing business with that company or the Customers that walks out of the one door in a given period. This Includes customers who moved over to the competition and the one who has terminated the services or stopped purchasing the product from the same buyer. Customer Churn, Is the Serious problem for a Flourishing as well as established Business.

**Problem Statement**

Churn prediction means detecting which customers are likely to leave a service or to cancel a subscription to a service. It is a critical prediction for many businesses because acquiring new customers or clients are often costlier more than retaining the existing ones.Businesses are very keen on measuring churn because keeping an existing customer is far less expensive than acquiring a new customer. New business involves working leads through a sales funnel, using marketing and sales budgets to gain additional customers.

Existing customers will often have a higher volume of service consumption and can generate additional customer referrals.Preventing customer churn is critically important to the telecommunications sector, as the barriers to entry for switching services are so low.

**Data Analysis**

* For the Data Analysis I have loaded few libraries and I have imported the Data Set as df1 which was in csv format. To keep original file Safe and Intact, I have loaded the same data set in Data Frame as instance named df.
* As second step I used the pd.set\_option to display all columns and all rows present, Later I loaded top 3 and bottom 3 rows to understand the pattern. We can also use samples to check random contents.
* I have Checked the data type, shape, size and dimension of the rows. I used info check to the contents. If they are Categorical Based, Integer Based or float Based. And found 01 Float, 02 integer, 18 object Based Columns. Here I also found Total charges columns to be object based which should be integer or float type.
* Third Step I explained the columns.
* Forth step: Null values and visualization of the Null Values, by this Method I found All content to be intact.
* Fifth Step: I checked the Number of Unique Values and contents Uniqueness of the for all Categorical, Numerical and Float Column.
* Sixth Step: I found the Empty rows in Total Charges Column and by replace method I Imputed them with np.nan Values and converted the rows to the float type and later I replaced those by fillna method by mean values.
* Seventh Step: Checked the duplicated rows and found No Duplicated Contents.
* Eight Step: I dropped those columns which Won’t add much importance for prediction and Finally I checked all Rows after imputation and Presented a Visualization for the same.

**Exploratory Data Analysis.**

##### **I Have used Mainly Seaborn Library along with Matplot library for visualization. First, we have used univariate, Bivariate, Multivariate graphs.**

**1. Univariate Analysis (Observation)**

1. Total Gender Columns Shows.

Out of Total 7043 Observations, Male Counts 3555 and Female Counts 3488.

Hence Male User counts are more than Female.

1. Senior Citizen Counts

Out of Total Customers, Graph Shows 5901 are young whereas 1142 shows Senior Citizen.

Maximum Young Customers.

1. Partner Counts.

Out of Total Customers, Graph Shows 3641 are No partner and 3402 are with Partner.

Hence Maximum Visitor are Without Partners

1. Tenure (Shows Top five Tenure Value\_counts)

* 01-613
* 72-362
* 02-238
* 03-200
* 04-176

Maximum Counts lies between 0 to 5 followed by 65 to 70 Here 0 to 5 very New whereas 65 to 70 is old customers.

1. Phone Services Counts Shows:

The One using Phone Services are 6361 and the one non user of Phone Service is 682

### 

### Graph Shows Maximum User and Non-Users are Female.

1. Multiple Lines Counts Shows:

* Non-User= 3390,
* Multiple Line User is 2971,
* No Phone Service User is 682

Maximum users are non-user of Multiple Lines. Maximum Male uses Multiple lines.

1. Internet Service Users Counts

* Fiber optic 3096
* DSL 2421
* No 1526

Maximum user of Fiber Optic followed by DSL, Both Male and Female Uses Internet Services

1. User Using Online Security Shows

* Non-User of Online Security 3498
* User 2019
* No internet User 1526

Maximum are non-User of online Security; Maximum Female User are Online Security User.

1. Online Backup Counts

* Non-User 3088
* User Using Online Backup 2429
* No internet User 1526

Maximum User are with No Online Backup, Maximum Female uses Online Backup

1. Device Protection Shows.

* Non-User of Device Protects 3095
* User of Device Protects 2422
* No internet Service User 1526

### Maximum user with No and Yes Device protection is Male.

1. Tech Support

* Non-Tech Support Users 3473
* Tech Support Users 2044
* No internet service 1526

#### Male and Female User Both Uses Tech Support, Uses of Tech Support More by Female

1. Streaming Movies Users Counts

* Non-Streaming Movies User 2785
* Streaming Movies User 2732
* No internet service User 1526

### Maximum Streaming Movies are used by Male as well as Female, Maximum NONE user is Male

1. Contact Shows

* Month-to-month Contract Users 3875
* Two-Year Contract Users 1695
* One year Contract Users 1473

#### Maximum One-year and Month to Month User are Male, Two Years contact are maximum by Male

1. Paperless Billing

* Paperless Billing Users 4171
* No Paperless Billing Users 2872

#### Both Male and Female Uses Paperless billing, No Paperless Billing User are maximum male

1. Total Charges.

* Per Person Sale Stands 2283.3004408418697

#### Maximum User paying total Charges are between is approx. 100, Maximum Buyers pays Total Charges 0 to 2000.

**EDA conclusion Remarks:**

* Company is Making a loss of 27 percentage due to churn that amounts per 1,39,131 Per Month
* Mainly Non-User of Services are showing High churn.
* Maximum Senior citizen have high churn.
* Need to introduce more attractive offers non partner customers.
* Phone service, Multiple Lines, Fiber Optic users and have high churn Rate, needs to focus on those Customers and Customers feedbacks to be taken by deploying Tech support as many users don’t use or might not be aware of the Tech Support.
* Try Selling long term Packages to those customers as Tenure increases churn goes down.
* Non-Security Users, Non-Back up Users, Non-device protection users, and Non-Tech Support users have maximum churn.
* Stream tv, Stream Movies users have high Churn. Quality of Services Received by them might not be up to the mark so Maximum Churn can be Seen among these Customers.
* Month to Month Contact customers needs extra attention as change of churn is very high.
* Paperless Billing, Electronic Check, Month- to-Month Bill paying Customers opt for these options but Graph also shows that maximum churn.
* Payment Method shows Electronic Check has maximum preference as compare to other payment Methods, Graph Shows highest churn on Electronic Check Payments users. Company Needs to influence their make payments by other modes by giving some offers or some free benefits that would make the loss narrow.
* Try Selling to Long Term packages to the customers to avoid customer churn.
* Make customers aware of the services that is being provided by the company.
* Dependents has less churn as compare to non-Dependents. Try to provide the customers with more efficient Tech Support so that, Churn can be drawn to least.

# **VARIOUS TOOLS FOR THE ANALYSIS**

## **Statistical graph (Descriptive Graph)**

Descriptive graph gives the idea of Data Distribution like mean, Median, standard Deviation, minimum maximum etc.

* Shows maximum outliers in total charges columns.
* All values to be intact

## **Correlation Graph**

## This GRAPH shows that correlation with the Target Variable.

* Contact is very negatively correlated with the Target variable.
* Monthly Charge is very positively related with Target variable.

## **Outliers**

* To Check Outliers, I have used Boxplot.
* Skewness
* To check the skewness, I have used distplot.

**DATA PREPROCESSING:**

**DATA-CLEANSING**

* Step 1. Correlation:

The columns which was highly negative was removed along with the columns that were showing the shows zero relations with the target variable were dropped.

* Step 2. Zscore:

Zscore method was used to remove the outliers from the columns. After removing Outliers, the data set is divided into two-part train and test data set

* Step 3. Skewness:

Skewness was removed by using Power Transformer and after using Skewness I found the skewness to be more than +/-65. However, was categorical, hence dropped.

* Step 4. Standardization:

Standard Scaler to bring the data under one scale.

* Step 5. Principle Component Analysis

PCA was working worse to the model, Hence Bypassed the PCA process.

**SMOTE**

**Before sending the data for Training and Testing, one need to check the dataset Shape of the Target Variable. While checking the data set shape, the data set was found to be imbalanced, So SMOTE technique was used to Overcome such problem.**

**Model Training: Methods Used are**

1. Importing import Library for Model Training.
2. However, In Data Set the Target variable had two categorical variables, 0 and 1 so Logistic Regression was used, If the Target variable would have been more than 3 then the classification method can be used in those cases.
3. The Models and Algorithm used for Training the Data Set are as follows.
   1. Logistic Regression
   2. Decision Tree Classifier
   3. KNeighbors Classifier
   4. SVM Classifier
   5. Gaussian NB
   6. Stochastic Gradient Descent Classifier (Linear Model)

Ensemble Technique Models Include

* + 1. Extra Trees Classifier
    2. Random Forest Classifier

Boosting Technique Models Include.

* + - Ada Boost Classifier
    - Gradient Boosting Classifier

Neural Network Model Include.

* MLP Classifier

1. **Metrics Used for checking errors:**
   * Accuracy score
   * Confusion Matrix
   * And Classification Report were Used to find the errors.

These Models (Algorithms) were used to find the best fit models. After Using the algorithm, The Training and Testing errors were checked by using Metrics. The cross-validation Score were checked for all those models. After the completion of Cross-Validation.

The model AUC\_ROC\_Graph were created to check for all those models’ prediction capabilities. And cross Validation Score was Noted. However, Finalization of the model for hyper-parameter Tuning was done by checking the Model Metrics like accuracy score, Training Score and the most import is the CV\_Score.

From these Model Selection Procedure, the model with better cv-score and CV\_Score near to training score were selected, those were used for Hyper-Parameter Tuning.

**HYPER-PARAMETER TUNING**

Each model has its own sets of parameters that need to be tuned to get optimal output. For every model, our goal is to minimize the error or say to have predictions as close as possible to actual values. This is one of the cores or say the major objective of hyperparameter tuning.

**Steps Used for WORKING:**

For the Hyper parameter tuning the main Library that is used is Grid Search CV other models like Random Search and Manual Search can be used by the user as per their system configuration or Requirement**.** Here the parameters of the certain (Selected) models need to be passed to the Grid Search CV, in the form of a dictionary and then the Grid Search CV finds the best parameter for the given Model and gives the score of the parameters.

These parameters are again passed to the given Algorithms and the score is being checked for the models and Cross Fold validation score is being noted again with testing score and then the model is being saved.

Before Saving the models Both the Models Needs to be Compared on the Basis of CV\_Score and Testing Score. The one with the Best close by and high CV\_Score and Testing Score is being saved.

## **MODEL SAVING**

The Two Main Library that can be used to save the model is Job lib and Pickle library.

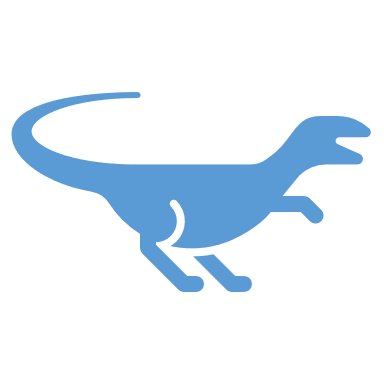
* First Step is to import the library.
* By using Dump method, the model is being saved along with the model Name.
* Model is Being Loaded back and score is being checked.
* Prediction the Model is being checked with the predicted and original.

**CONCLUSION:**

While doing the EDA I found very, we got to know that A crucial part of running a business is when one is able to understand the value of your customers. This is even more important in the digital age where it is easier than ever to get the attention of your customers. Churn helps you understand how many customers are leaving your business, and why they’re leaving. This is essential for understanding whether you’re marketing and Customer Retention strategies are working. When a business loses a customer, it’s not just the revenue from that customer that’s lost. There are additional costs associated with attracting new Customers, which include marketing spending and the cost of sales. We got to know more about the loss that is involved during a customer churn. If this continues, how would it impact on the company month to month basis. Customer are somewhere unawareness of services provided by the company, or they are not using the services that is provided by the company, Customers migrate with one to other company for a better service.

Churn is an import Phenomena for customer where customer will always try to be find the best opportunities for them. We can’t stop customer Churn but we can reduce this to maximum Level by following strategies.

* Company Should Upgrade to Modern Technology
* Providing Better Customer Service or Support Team.
* Easy Interaction with Customer.
* Providing Alternative Modes.
* Identifying the Needs the customers.
* Improving Customer Retention.
* Giving Free Benefits or Discount Coupons to Customer to prevent churn (Reworking on price Models)
* Selling them long term Plans
* Providing Customer Feedback Forms for Each Requests.



**The End**