The University of Texas at Dallas

Halfway Report

Project Title – Customer Churn Analysis for a Telecom Company

Professor – Mr. Zhe Zhang

Group Members

Name	NetID
Akshay Manchekar	anm230012
Kaustubh Darange	kxd230024
Siddhi Hulwane	sxh240015
Ritika Namdeo	rxn24006

Introduction

In the telecom industry, keeping customers is essential for success. Companies spend a lot of money to attract new customers, but if many of those customers leave, it can hurt profits. Analyzing customer churn—the rate at which customers leave—helps telecom companies understand why people are leaving and how they can keep them. By identifying patterns and behaviors linked to customers leaving, companies can take action to improve service, keep customers satisfied, and reduce the number of customers who switch to other providers. This focus on customer retention not only increases profits but also builds stronger customer relationships and helps the company compete in the market.

Objective

The main goal of this project is to create a model that predicts which customers are likely to leave. By looking at how customers use services, what plans they have, and how often they contact support, this project aims to uncover the main reasons for churn. Understanding these reasons will help the company create specific actions to keep at-risk customers from leaving. The end goal is to provide insights that the company can use to reduce churn, make better use of resources, and improve customer loyalty.

Data Description

This dataset is from a telecom company and is specifically designed for churn analysis. It contains customer information and usage data that will help us analyze patterns related to customer churn.

The dataset has 3,334 entries (or records), each representing a customer. It includes 11 columns, each containing specific information about the customer.

Below are some of the main variables we'll be focusing on:

- **Churn**: Indicates if the customer left the company or stayed.
- **AccountWeeks**: Number of weeks the customer has had an account.
- **ContractRenewal**: Shows whether the customer recently renewed their contract.
- **DataPlan**: Whether the customer has a data plan.
- **DataUsage**: Amount of data used by the customer.
- **CustServCalls**: Number of times the customer contacted customer service.
- **DayMins**: Total minutes used during the day.
- **DayCalls**: Number of calls made during the day.
- **MonthlyCharge**: The customer's monthly charge.
- **OverageFee**: Charges for usage over the customer's data plan.
- **RoamMins**: Minutes spent on roaming.

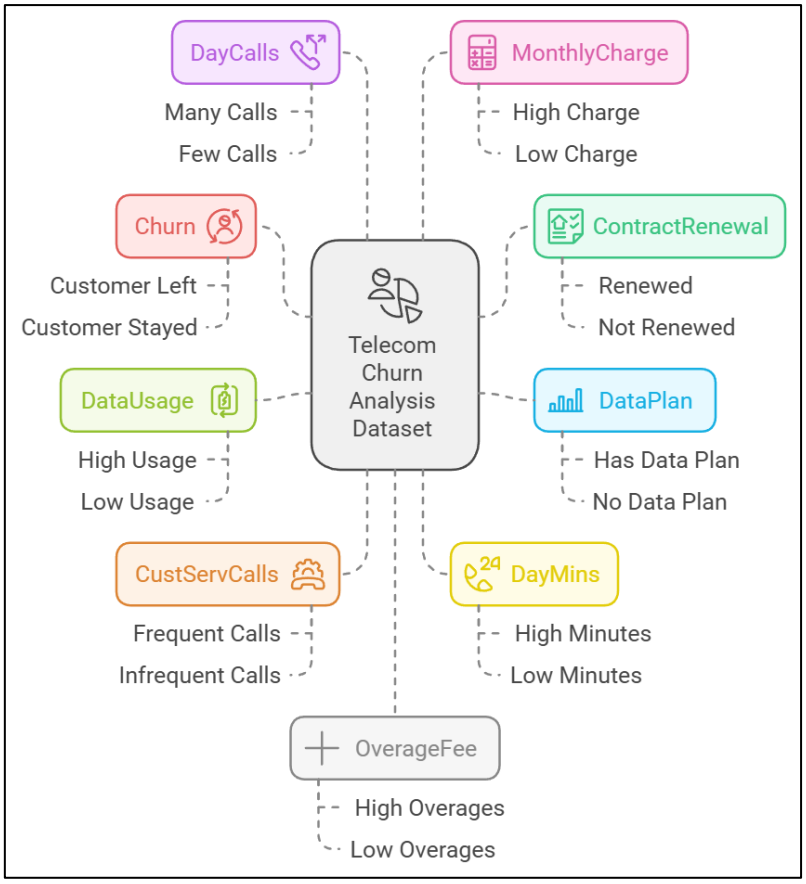


Fig. 1: Dataset Attributes

The dataset is submitted as a separate file along with this report.

Steps of the Project

1. Data Preprocessing

- ***Loading the Dataset***: We used R to load the dataset from a CSV file, preparing it for analysis.
- ***Handling Missing Data***: Checked for any missing or null values in the dataset and handled them appropriately, using methods like mean or median imputation where needed.
- ***Data Type Conversion***: Converted categorical variables, such as Churn, ContractRenewal, and DataPlan, into factors for better handling in R.
- ***Data Normalization***: Normalized continuous features like DataUsage, MonthlyCharge, and OverageFee where necessary to ensure a consistent scale across the dataset.

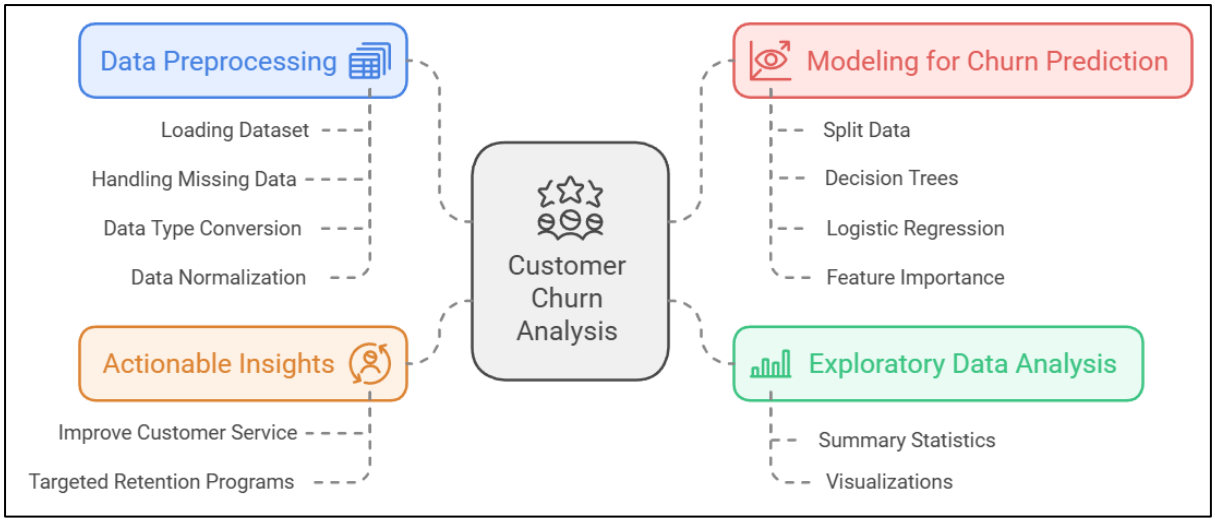


Fig. 2: Steps of the project

2. Exploratory Data Analysis (EDA)

- ***Summary Statistics:*** Calculated basic summary statistics (mean, median, minimum, maximum) for all columns to get a clear view of data distribution.

- ***Visualizations:***
  - Bar Chart of Churn Rate: Created bar charts to show churn rates across categories like ContractRenewal and DataPlan.
  - Correlation Heatmap: Visualized correlations between continuous variables such as DataUsage, CustServCalls, and MonthlyCharge.
  - Boxplots: Used boxplots to explore distributions of numeric features like MonthlyCharge and OverageFee across customers who churned and those who didn't.

3. Modeling for Churn Prediction

- ***Split Data***: We will divide the data into training and testing sets (using an 80/20 split) to train and evaluate models.
- ***Model 1 - Decision Trees***: We will train a decision tree model to predict churn, using evaluation metrics like accuracy, precision, and recall.
- ***Model 2 - Logistic Regression***: We will build a logistic regression model and assess its performance using similar metrics.
- ***Feature Importance***: We will analyze feature importance from each model to identify key factors that contribute to customer churn.

4. Actionable Insights

Based on the model results, we will identify strategies for reducing churn, such as improving customer service response times and introducing targeted retention programs for high-churn customer segments.

Summary Statistics (in brief)

```
# View the first few rows of the dataset to verify the data is loaded correctly
head(telecom_data)
```

	Churn	AccountWeeks	ContractRenewal	DataPlan	DataUsage	CustServCalls	DayMins	DayCalls	MonthlyCharge	OverageFee	RoamMins
1	0	128	1	1	2.7	1	265.1	110	89	9.87	10.0
2	0	107	1	1	3.7	1	161.6	123	82	9.78	13.7
3	0	137	1	0	0.0	0	243.4	114	52	6.06	12.2
4	0	84	0	0	0.0	2	299.4	71	57	3.10	6.6
5	0	75	0	0	0.0	3	166.7	113	41	7.42	10.1
6	0	118	0	0	0.0	0	223.4	98	57	11.03	6.3

```
# Check the structure of the dataset
str(telecom_data)
```

```
'data.frame':   3333 obs. of  11 variables:
 $ Churn          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AccountWeeks   : int  128 107 137 84 75 118 121 147 117 141 ...
 $ ContractRenewal: int  1 1 1 0 0 0 1 0 1 0 ...
 $ DataPlan       : int  1 1 0 0 0 0 1 0 0 1 ...
 $ DataUsage      : num  2.7 3.7 0 0 0 0 2.03 0 0.19 3.02 ...
 $ CustServCalls  : int  1 1 0 2 3 0 3 0 1 0 ...
 $ DayMins        : num  265 162 243 299 167 ...
 $ DayCalls       : int  110 123 114 71 113 98 88 79 97 84 ...
 $ MonthlyCharge  : num  89 82 52 57 41 57 87.3 36 63.9 93.2 ...
 $ OverageFee     : num  9.87 9.78 6.06 3.1 7.42 ...
 $ RoamMins       : num  10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...
```

```
# Check the summary of the data
summary(telecom_data)
```

```
     Churn         AccountWeeks   ContractRenewal     DataPlan        DataUsage      CustServCalls
 Min.   :0.0000   Min.   :  1.0   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
 1st Qu.:0.0000   1st Qu.: 74.0   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.000
 Median :0.0000   Median :101.0   Median :1.0000   Median :0.0000   Median :0.0000   Median :1.000
 Mean   :0.1449   Mean   :101.1   Mean   :0.9031   Mean   :0.2766   Mean   :0.8165   Mean   :1.563
 3rd Qu.:0.0000   3rd Qu.:127.0   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.7800   3rd Qu.:2.000
 Max.   :1.0000   Max.   :243.0   Max.   :1.0000   Max.   :1.0000   Max.   :5.4000   Max.   :9.000
    DayMins          DayCalls     MonthlyCharge      OverageFee       RoamMins
 Min.   :  0.0   Min.   :  0.0   Min.   : 14.00   Min.   : 0.00   Min.   : 0.00
 1st Qu.:143.7   1st Qu.: 87.0   1st Qu.: 45.00   1st Qu.: 8.33   1st Qu.: 8.50
 Median :179.4   Median :101.0   Median : 53.50   Median :10.07   Median :10.30
 Mean   :179.8   Mean   :100.4   Mean   : 56.31   Mean   :10.05   Mean   :10.24
 3rd Qu.:216.4   3rd Qu.:114.0   3rd Qu.: 66.20   3rd Qu.:11.77   3rd Qu.:12.10
 Max.   :350.8   Max.   :165.0   Max.   :111.30   Max.   :18.19   Max.   :20.00
```

```
# Check for missing values
sum(is.na(telecom_data))
```

```
[1] 0
```

Calculating mean for Monthly Charge

```
> mean(telecom_data$MonthlyCharge)
[1] 56.30516
```

Calculating median for Monthly Charge

```
> median(telecom_data$MonthlyCharge)
[1] 53.5
```

Calculating standard deviation for Monthly Charge

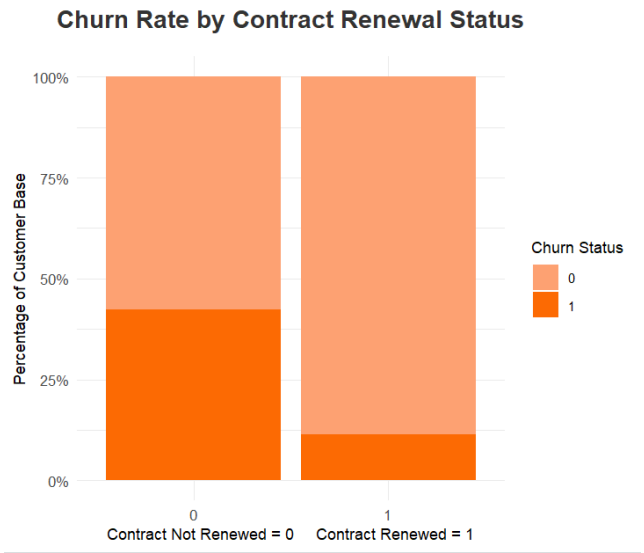
```
> sd(telecom_data$MonthlyCharge)
[1] 16.42603
```

Calculating variance for monthly Charge

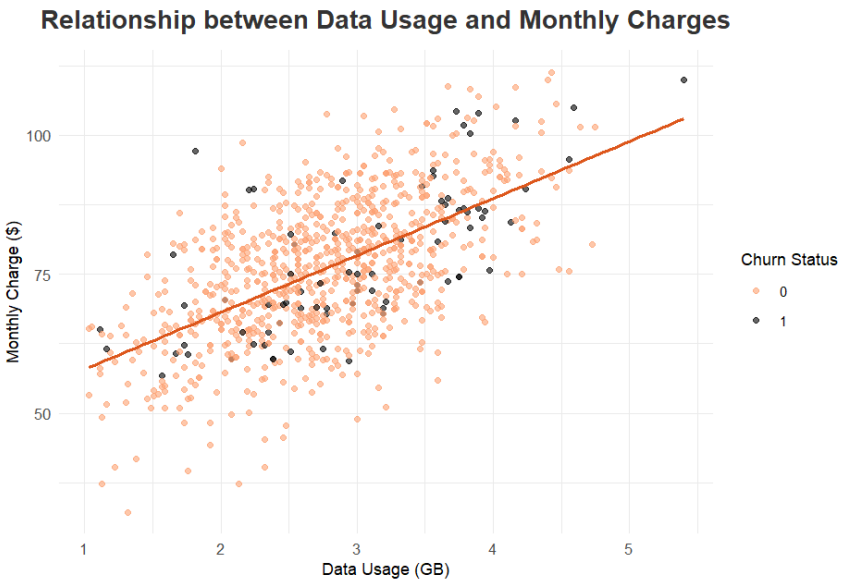
```
> var(telecom_data$MonthlyCharge)
[1] 269.8145
```

Key Visualizations & Rough Findings (in brief)

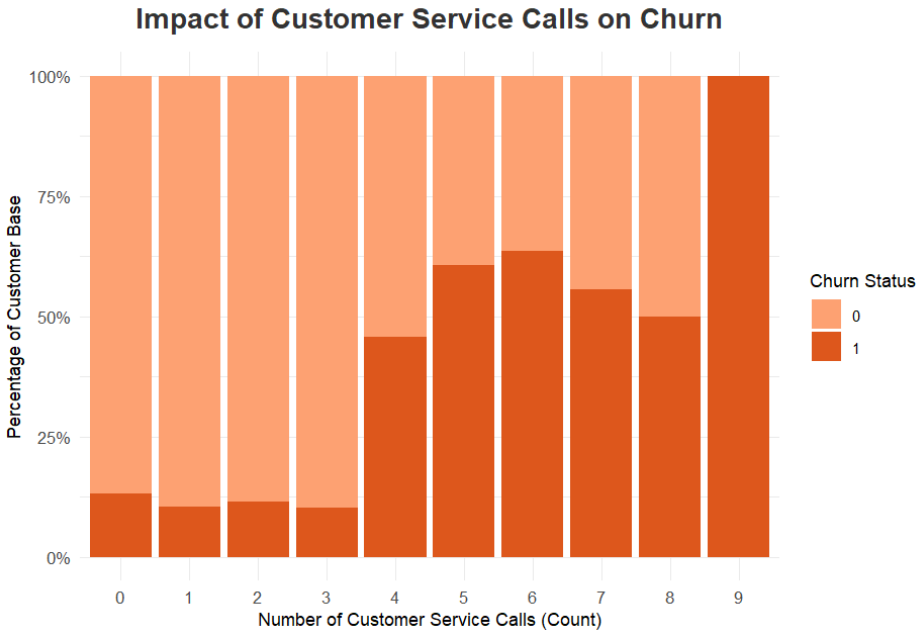
The customers whose contract got renewed had low churn rate.



Majority of the customer base is using data between 2-4 GB per month and have a monthly charge of around $75.



Customers who received high number of service calls had the highest churn rate.



Churn Rate

- Total customers = 3333
- Churned customer = 483
- Churn Rate = 14.49%

Data Mining

Association Rules: Apriori Algorithm

- Minimum Support = 40%
- Minimum Confidence = 80%

```
    lhs                                          rhs                  support   confidence coverage  lift      count
[1] {}                                        => {Churn=0}            0.8550855 0.8550855  1.0000000 1.0000000 2850
[2] {}                                        => {ContractRenewal=1} 0.9030903 0.9030903  1.0000000 1.0000000 3010
[3] {DataPlan=Basic Plan}                     => {Churn=0}            0.6024602 0.8328494  0.7233723 0.9739955 2008
[4] {DataPlan=Basic Plan}                     => {ContractRenewal=1} 0.6540654 0.9041891  0.7233723 1.0012167 2180
[5] {Churn=0}                                 => {ContractRenewal=1} 0.7992799 0.9347368  0.8550855 1.0350425 2664
[6] {ContractRenewal=1}                       => {Churn=0}            0.7992799 0.8850498  0.9030903 1.0350425 2664
[7] {DataPlan=Basic Plan, Churn=0}            => {ContractRenewal=1} 0.5634563 0.9352590  0.6024602 1.0356206 1878
[8] {ContractRenewal=1, DataPlan=Basic Plan} => {Churn=0}            0.5634563 0.8614679  0.6540654 1.0074640 1878
```

Summary of 8 rules

- Rules 1 and 2 show that a significant majority of customers do not churn and renew their contracts.
- Rules 3 and 4 suggest that Basic Plan customers tend to have high renewal rates and lower churn.
- Rules 5 and 6 reinforce that not churning customers are highly likely to renew contracts and vice versa.
- Rules 7 and 8 further highlight the strong relationship between having a Basic Plan, not churning, and renewing contracts.