Telecom Churn Case Study

by- Tilak Shah

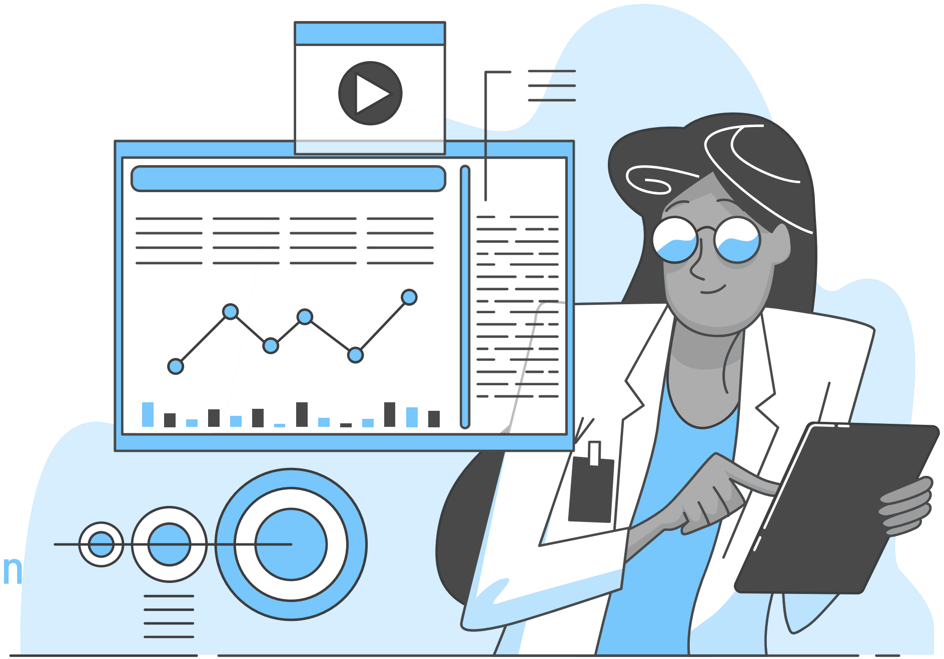
Table of Contents

01 Problem statement

02 Problem approach

03 EDA Correlations

04 Observations / Conclusion



Problem Statment

1. **High Churn Rates**: Annual churn rate of 15-25% in the telecom industry.
2. **Cost Implications**: Acquiring new customers costs 5-10 times more than retaining existing ones.
3. **Business Goal**: Retaining high-value, profitable customers is critical.
4. **Challenge**: Need to predict which customers are at high risk of churn.
5. **Focus Area**: Prepaid customers in the Indian and Southeast Asian markets.
6. **Churn Definition**: Usage-based churn - customers with zero usage (calls, internet, SMS) over a specific period.
7. **Revenue Impact**: 80% of revenue comes from the top 20% of customers. Reducing churn among these high-value customers is essential.

Business Objective

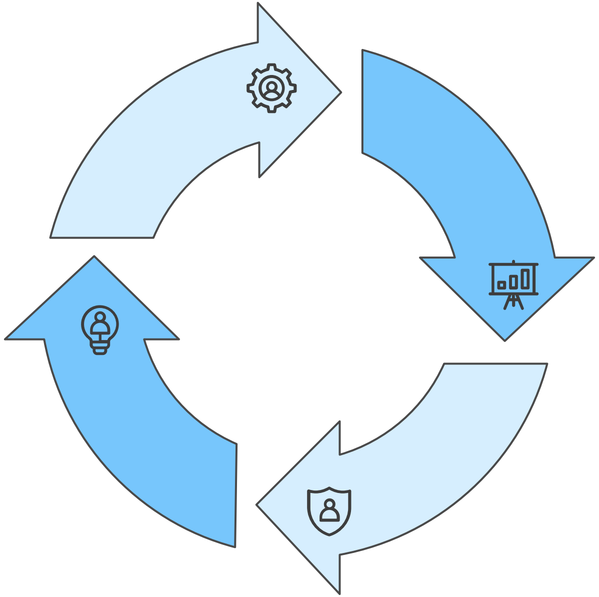
1. **Predictive Modeling**: Develop models to predict customers at high risk of churn.
2. **Identify Indicators**: Determine the main indicators of churn, such as usage patterns and recharge frequency.
3. **Retention Strategies**: Create effective strategies to retain high-value customers.
4. **Proactive Measures**: Implement early intervention tactics based on predictive insights.
5. **Continuous Monitoring**: Establish ongoing monitoring systems for early detection of churn risk.
6. **Personalized Engagement**: Offer personalized incentives and services to retain high-value customers.

Approach

1. Data Understanding
2. Data Cleaning and Handeling Missing Values
3. Filtering High Value Customer
4. Deriving Churn
5. Deriving New Features
6. Data Visualization-Univariate Analysis/Bivariate Analysis
7. Data Preparation
8. Data Modeling and Eavlaution
9. Conclusions

Data Cleaning

1. **Impute Recharge Columns**:Recharge Columns with more than 70% missing values imputed with 0 (no recharge done; minimum observed value was 1).
2. **Drop Identifier Columns**:Dropped mobile_number and circle_id as they are not useful for analysis.
3. **Drop Columns** : Droping Columns with more the 70% missing vales.
4. **Handle Categorical Missing Values**:Replaced missing values in categorical columns with '-1', introducing it as a new category.
5. **Impute Numerical Missing Values**:Imputed remaining missing values in numerical columns with the median.



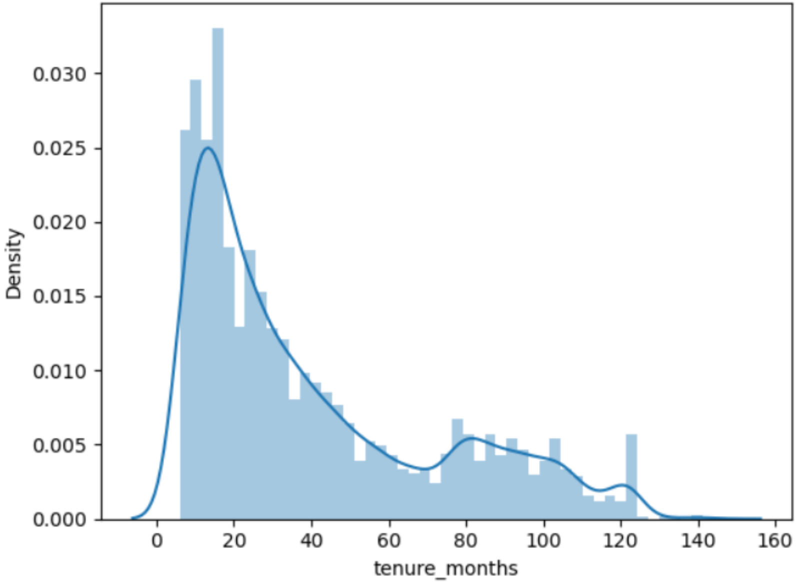
Filter high-value customers

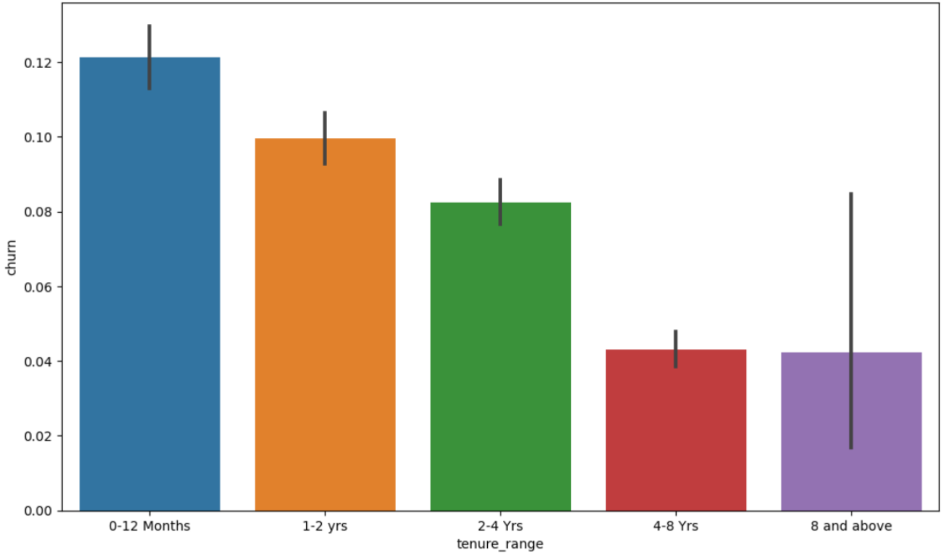
1. **Calculate Total Data and Call Recharge Amounts** for June and July.
2. **Determine the 70th Percentile** of the average recharge amount for June and July.
3. **Filter Customers** by retaining only those with recharge amounts above the 70th percentile.
4. **Result**: 30,001 customers selected for further analysis.

Derive churn

1. **Data Source**: Use data from the 9th month to create the churn variable.
2. **Attributes Used**: Analyze specific customer behaviors like incoming and outgoing calls, and data usage.
3. **Churn Determination:** If a customer shows no activity in terms of incoming and outgoing calls, as well as data usage, mark them as churned (1); otherwise, mark as not churned (0).

EDA

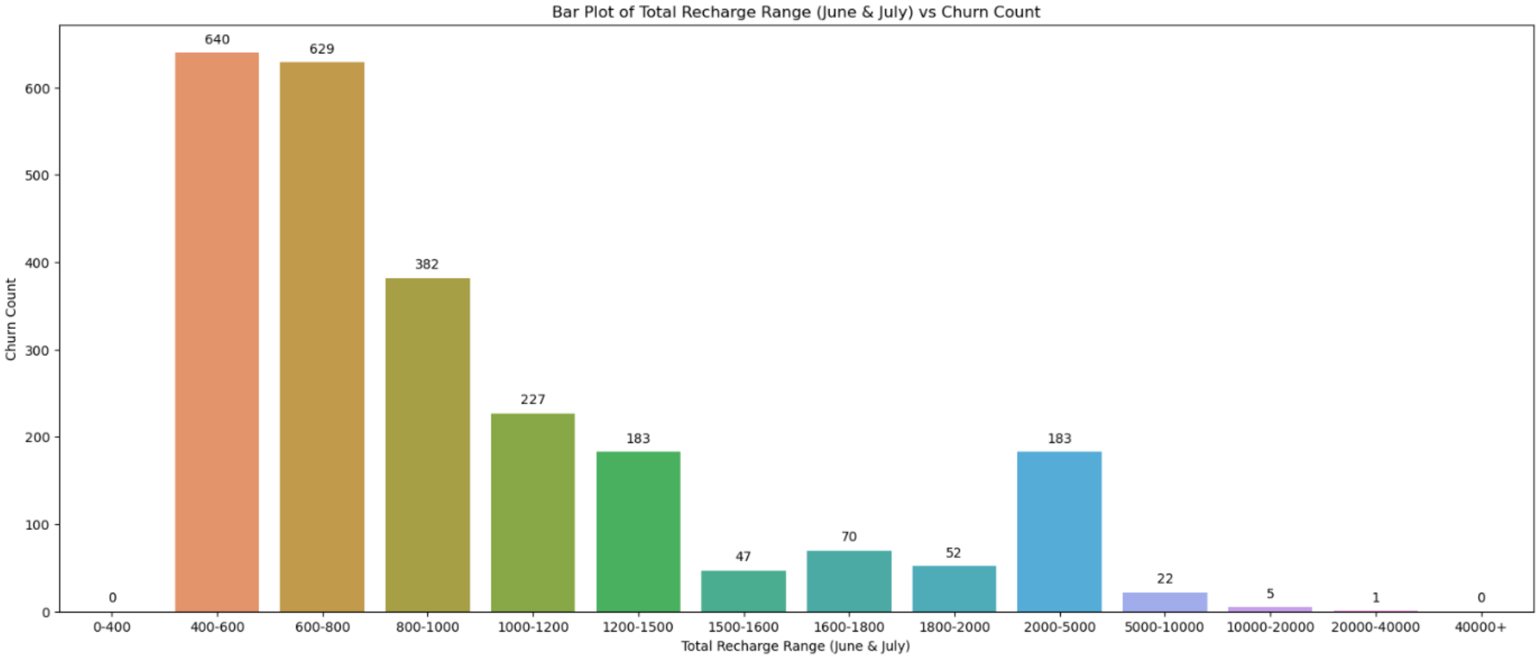




We can see that we have more customers with tenure 20 to 40 months.

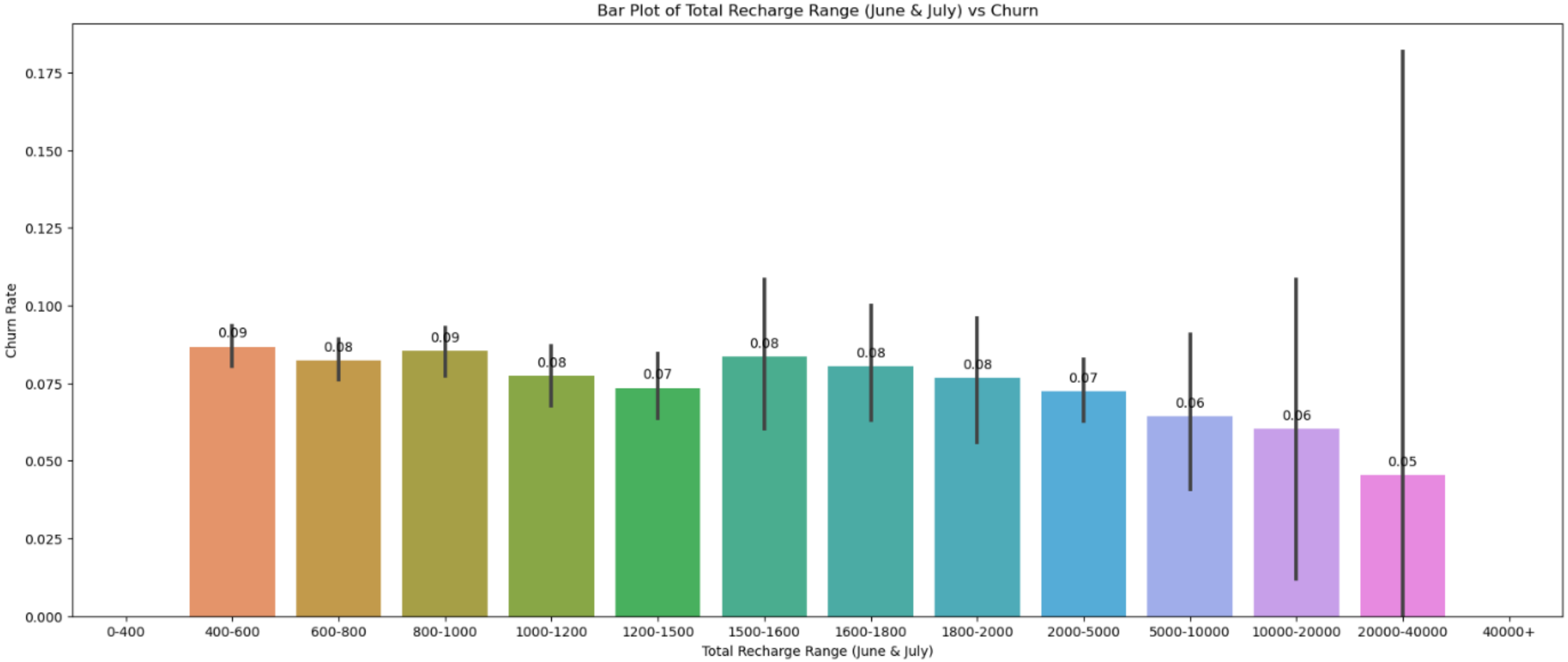
We can observe that the highest churn rate occurs within the initial 0-12 months, gradually decreasing as customers remain on the network.

EDA



We can see the highest number of Churns in lower Range. (Lower Range of Total Recharge for the month of June and July

EDA



We observe that all recharge ranges exhibit similar churn rate proportions. However, customers with a recharge amount in the range of 400 to 600 and 800 to 1000 tend to churn slightly more than others, while those with a recharge amount above 5000 tend to have little less Churn Rate.

Observation & Conclusion

Observation:
Logistic Regression
. Train Accuracy : ~79% . Test Accuracy : ~80%
Logistic regression with PCA
. Train Accuracy : ~91% . Test Accuracy : ~92%
Decision Tree with PCA:
. Train Accuracy : ~93% . Test Accuracy : ~92%
Random Forest with PCA:
. Train Accuracy :~ 91% . Test Accuracy :~ 92%

Conclusion:
Based on accuracy, the Random Forest model is the most effective for predicting churn as their Test and Train Accuracy are close to each other
Outgoing calls, whether local (same operator mobile, other operator mobile, fixed lines), STD, or special, are crucial in assessing the likelihood of churn. Therefore, the operator should focus on analyzing Outgoing call data and consider offering special incentives to customers with decreasing Outgoing call volumes.

Thank You

Tilak Shah