Telecom Churn Case Study

Analysis Approach :

The telecommunications industry experiences an annual churn rate of 15-25%, and with the cost of acquiring a new customer being 5-10 times higher than retaining an existing one, customer retention has become more critical than customer acquisition. In this case study, we analyze customer usage data over a period of four months to predict customer churn for a leading telecom firm. The focus is on identifying customers at high risk of churn and determining the key indicators of churn.

Two approaches are considered for predicting churn: usage-based churn and revenue-based churn. In this study, we focus solely on usage-based churn, defined as customers with zero usage (in terms of calls, internet, etc.) over a specific period. Given that approximately 80% of revenue in the Indian and Southeast Asian markets comes from the top 20% of customers (high-value customers), this study specifically targets reducing churn among these high-value customers to minimize revenue leakage.

The dataset includes customer-level information for four consecutive months: June, July, August, and September, encoded as 6, 7, 8, and 9, respectively. The goal is to predict churn in the ninth month using features from the first three months. This is a classification problem where the objective is to predict whether a customer is likely to churn. Various models have been applied, including Baseline Logistic Regression, Logistic Regression with PCA, PCA combined with Random Forest, and PCA combined with XGBoost.

Analysis Steps

Data Cleaning and Exploratory Data Analysis (EDA):

1. **Package Import and Dataset Loading:** We started by importing the necessary packages and libraries, followed by loading the dataset into a DataFrame.
2. **Initial Data Inspection:** We checked the number of columns, their data types, the count of null values, and the distribution of unique values to understand the data better and ensure the columns were correctly typed.
3. **Duplicate Check:** We checked for duplicate records (rows) and found none.
4. **Indexing:** The column 'mobile_number' was identified as a unique identifier and set as the index to retain customer identity.
5. **Column Renaming:** Some columns did not follow a consistent naming convention, so we renamed them to ensure uniformity.
6. **Data Type Conversion:** Columns with 29 or fewer unique values were treated as categorical, while the rest were treated as continuous. Date columns were converted from 'object' to a proper datetime format.
7. **Filtering for High-Value Customers (HVC):** Since the focus is on HVCs, we filtered customers whose 'Average_rech_amt' in months 6 and 7 was greater than or equal to the 70th percentile, categorizing them as HVCs.
8. **Missing Values Handling:**
    - Columns with over 50% missing values were dropped.
    - We conducted a month-wise analysis of missing values, given the independence of each month's data.
    - For columns with similar missing value patterns, we considered imputing missing values with zeros.
    - Missing values in the 'last_date_of_the_month' column were imputed based on the respective month.
9. **Dropping Redundant Columns:** Columns with only one unique value were dropped as they provided no analytical value.

10. **Target Variable Tagging:** The churn variable (our target) was tagged. After imputations, we dropped columns from the ninth month (the churn phase).
11. **Final Data Structure:** After all processing, we retained 30,011 rows and 126 columns.

Exploratory Data Analysis (EDA):

1. **Revenue Patterns:** Many users showed negative average revenues in both periods, indicating a higher likelihood of churn.
2. **Customer Preferences:** Most customers preferred plans in the '0' category.
3. **Customer Longevity (AON):** Customers with shorter account lifetimes ('aon') were more likely to churn compared to those with longer durations.
4. **Revenue Stability:** Customers who were about to churn exhibited unstable revenue patterns.
5. **Usage Patterns:**
    - Customers whose ARPU (Average Revenue Per User) decreased in the 7th month were more likely to churn.
    - A significant drop in total outgoing minutes (total_og_mou) from month 6 to 7 was a strong churn indicator.
    - Similar trends were observed with incoming minutes (total_ic_mou) and data usage (2G and 3G).
6. **Correlation Analysis:** Conducted to understand relationships between variables.
7. **Derived Variables and Outlier Treatment:** We created derived variables and removed the original variables used in their creation. Outliers were treated by capping them at the 99th percentile.
8. **Categorical Variable Grouping:** Classes with minimal contribution in categorical variables were grouped into an 'Others' category.
9. **Dummy Variable Creation:** Created for categorical variables.

Pre-processing Steps:

1. **Train-Test Split:** The dataset was split into training and testing sets.

2. **Class Imbalance:** The data exhibited a high class imbalance with a ratio of 0.095 (class 1: class 0). The SMOTE technique was used to address this imbalance.
3. **Standardization:** Predictor columns were standardized to have a mean of 0 and a standard deviation of 1.

Modeling:

1. **Model 1: Logistic Regression with RFE & Manual Elimination (Interpretable Model):**
    - Key predictors of churn included lower local incoming calls from fixed lines, fewer recharges, and usage of 'monthly 2G/3G packages-0'.
2. **Model 2: PCA + Logistic Regression:**
    - **Train Performance:**
        - Accuracy: 0.627
        - Sensitivity: 0.918
        - Specificity: 0.599
        - Precision: 0.179
        - F1-score: 0.3
    - **Test Performance:**
        - Accuracy: 0.086
        - Sensitivity: 1.0
        - Specificity: 0.0
        - Precision: 0.086
        - F1-score: 0.158
3. **Model 3: PCA + Random Forest Classifier:**
    - **Train Performance:**
        - Accuracy: 0.882
        - Sensitivity: 0.816
        - Specificity: 0.888
        - Precision: 0.408
        - F1-score: 0.544
    - **Test Performance:**
        - Accuracy: 0.86
        - Sensitivity: 0.80

        - Specificity: 0.78
        - Precision: 0.37
        - F1-score: 0.51
4. **Model 4: PCA + XGBoost:**
    - **Train Performance:**
        - Accuracy: 0.873
        - Sensitivity: 0.887
        - Specificity: 0.872
        - Precision: 0.396
        - F1-score: 0.548
    - **Test Performance:**
        - Accuracy: 0.086
        - Sensitivity: 1.0
        - Specificity: 0.0
        - Precision: 0.086
        - F1-score: 0.158

Recommendations:

- **Focus on Customers with Lower Incoming Calls:** Customers with 1.27 standard deviations lower incoming calls from fixed lines are most likely to churn.
- **Monitor Recharge Frequency:** Users who recharge less frequently (1.2 standard deviations lower in the 8th month) are more likely to churn.
- **Model Selection:** Models with high sensitivity are ideal for churn prediction. The PCA + Logistic Regression model, with an ROC score of 0.87 and 100% test sensitivity, is recommended for predicting churn.