Telecom Churn Case Study

Analysis Approach

- The telecommunications industry experiences an average annual churn rate of 15 - 25%. Given the significant cost difference between acquiring new customers and retaining existing ones, customer retention has become a crucial focus.
- In this case study, we analyze customer-level data from a leading telecom firm, aiming to build predictive models to identify customers at high risk of churn and determine the main indicators of churn.
- Churn is predicted using two approaches: usage-based churn and revenue-based churn. We focus solely on usage-based churn for this case study.
- Approximately 80% of revenue in the Indian and southeast Asian markets comes from the top 20% of customers (high-value customers). Thus, reducing churn among high-value customers can significantly decrease revenue leakage. Consequently, this case study focuses on high-value customers exclusively.
- The dataset contains customer-level information for four consecutive months: June (6), July (7), August (8), and September (9).
- The business objective is to predict churn in the last month (ninth) using data from the first three months (June, July, and August).
- This is a classification problem where we aim to predict whether customers are likely to churn.

Analysis Steps

Data Cleaning and EDA

1. **Data Loading:** Import necessary packages and libraries and load the dataset into a dataframe.
2. **Data Understanding:** Check the number of columns, their data types, null count, and unique value count to gain insights and ensure correct data types.
3. **Duplicate Checking:** Check for duplicate records (rows) in the data.
4. **Indexing:** Set 'mobile_number' as the index to retain identity.
5. **Column Renaming:** Rename columns to ensure all variables follow the same naming convention.
6. **Data Type Conversion:** Convert columns into their respective data types, categorizing columns with less than or equal to 29 unique values as categorical and the rest as continuous.
7. **Date Columns Conversion:** Convert date columns to the proper datetime format.
8. **Filtering for High-Value Customers (HVC):** Filter for high-value customers based on the 'Average_rech_amt' of months 6 and 7, considering those whose average recharge amount is greater than or equal to the 70th percentile.
9. **Missing Values Handling:**
    - Check for missing values.
    - Drop columns with missing values greater than 50%.
    - Impute meaningful values for specific columns.
    - Drop columns with only one unique value.
10. **Target Variable Tagging:** Tag the churn variable (target variable).
11. **Drop Churn Phase Columns:** Drop churn phase columns (columns belonging to month 9).
12. **Retained Data Summary:** After data preparation, retain 30,011 rows and 126 columns.

Exploratory Data Analysis (EDA)

- Identify patterns and insights from the data.

- Analyze revenue, customer preferences, customer behavior, and other relevant factors.
- Perform correlation analysis and derive new variables.
- Perform outlier treatment and create dummy variables.

Pre-processing Steps

- Perform train-test split.
- Address class imbalance using SMOTE technique.
- Standardize predictor columns.

Modelling

1. **Model 1: Logistic Regression with RFE & Manual Elimination (Interpretable Model):**
   - Identify the most important predictors of churn.
2. **Model 2: PCA + Logistic Regression:**
   - Train and evaluate the model's performance.
3. **Model 3: PCA + Random Forest Classifier:**
   - Train and evaluate the model's performance.
4. **Model 4: PCA + XGBoost:**
   - Train and evaluate the model's performance.

Recommendations

Strongest Indicators of Churn

1. **Average Monthly Local Incoming Calls from Fixed Line:**
   - Customers who churn exhibit a lower average monthly local incoming calls from fixed line in the action period by 1.27 standard deviations compared to users who don't churn.
2. **Number of Recharges Done in Action Period:**
   - Customers who churn show a lower number of recharges done in the action period by 1.20 standard deviations compared to users who don't churn.
3. **Recharge Amount:**
   - Furthermore, customers who churn have done 0.6 standard deviations higher recharge than non-churn customers.
4. **Usage of Monthly 2G/3G Packages:**
   - Customers who churn are more likely to be users of 'monthly 2G package-0 / monthly 3G package-0' in action period.

Recommendations to the Telecom Company

1. Concentrate on users with lower than average incoming calls from fixed line.
2. Target users who recharge less frequently in the 8th month.
3. Utilize models with high sensitivity, such as PCA + Logistic Regression, for predicting churn.