Telecom Churn Case Study

By Poulami Saha, Navneet Mahar & Pradiptamay Banerjee

Batch : Data Science Program - December 2023

Summary of Data Preparation

1. **Data Import and Overview:**
   - Data was imported and inspected. It contains 99999 rows and 226 columns with various attributes related to customer usage and recharge information.
2. **Identifying High-Value Customers:**
   - High-value customers are defined as those whose average recharge amount in June and July is in the top 30% (70th percentile).
   - Created new columns to account for data recharge amounts.
   - Filtered data to include only high-value customers.
3. **Churn Label Assignment:**
   - Churn is defined as not using any calls or mobile internet in September.
   - Created aggregated columns for call and data usage in September.
   - Labeled churners based on whether these aggregated usage values are zero.
4. **Feature Engineering:**
   - Removed features that were not needed for the churn prediction.
   - Cleaned up data by removing columns with over 40% missing values.
   - Kept only relevant columns and further reduced the dataset to 143 columns.

Steps for Building Predictive Models

1. **Handling Missing Values:**
    - Your data now has fewer missing values. For the remaining missing values, consider:
        - Imputation (mean, median, or mode for numerical columns).
        - Using predictive imputation methods if missing values are substantial.
        - Removing rows with missing values if they are few and randomly distributed.
2. **Feature Selection:**
    - You may want to perform feature selection to further reduce dimensionality and focus on the most predictive features. Techniques include:
        - Correlation analysis to remove highly correlated features.
        - Feature importance from tree-based models (e.g., Random Forest).
        - Recursive Feature Elimination (RFE).
3. **Data Preprocessing:**
    - **Scaling:** Scale numerical features (e.g., using StandardScaler or MinMaxScaler) to ensure that all features contribute equally to the model.
    - **Encoding:** Convert categorical features into numerical formats if any remain.

Contd....

1. **Model Building:**
    - Split the data into training and testing sets.
    - Train several models and compare their performance. Suitable models for churn prediction include:
        - **Logistic Regression:** A good baseline for binary classification problems.
        - **Decision Trees / Random Forests:** Handle non-linearities and interactions well.
        - **Gradient Boosting Machines (e.g., XGBoost, LightGBM):** Often provide high performance for classification tasks.
        - **Support Vector Machines (SVM):** Effective in high-dimensional spaces.
2. **Model Evaluation:**
    - Use metrics such as Precision, Recall, F1-Score, ROC-AUC to evaluate model performance.
    - Cross-validation can help ensure that the model generalizes well to unseen data.
3. **Hyperparameter Tuning:**
    - Optimize model performance using techniques like Grid Search or Random Search to find the best hyperparameters.
4. **Model Interpretation and Deployment:**
    - Interpret the model to understand key features driving churn.
    - Deploy the model in a production environment to predict churn and help with customer retention strategies.

Introduction to the Problem

Objective:

- **Industry Context:** The telecom industry in India and South East Asia is highly competitive, with customers able to switch providers easily.
- **Main Goal:** Telecom operators aim to retain customers, especially high-value ones.

Challenge:

- **Need:** One of the leading telecom companies seeks to develop predictive models to identify customers at high risk of churn.

Data Overview

- **Dataset:** telecom_churn_data.csv
- **Initial Examination:** The dataset contains 226 columns, capturing various customer attributes and usage patterns.
- **Data Cleaning:** Handle missing values, incorrect entries.
- **Feature Engineering:** Identify relevant features for churn prediction.
- **Summary of the Work Done:**
- **Data Overview:**
- **Data Dimensions:** The dataset has 99,999 entries and 226 columns.
- **Data Types:** The dataset contains a mix of numeric (both float64 and int64) and object (string) data types.

Data Cleaning and Preparation:

- Created a copy of the original dataset.
- Removed columns related to recharge amounts and calculated the total data recharge amount for each month.
- Dropped the columns that were used to calculate total data recharge amounts.
- Computed the 70th percentile of the average recharge amount during the good phase (June and July) to define high-value customers.
- **Filtering High-Value Customers:**
- Filtered the dataset to include only high-value customers based on the 70th percentile of the average recharge amount.
- Verified the shape of the filtered dataset, which contains 30,001 entries.
- **Churn Labeling:**
- Determined which customers have churned by checking if they have not engaged in any calls or used mobile internet during the churn phase (September).
- **Feature Summation for Churn Analysis:**
- Created columns to sum up various metrics related to incoming and outgoing calls, as well as data usage for September.
- The next step likely involves analyzing these aggregated features to identify patterns or characteristics of churned customers.

Exploratory Data Analysis (EDA):

Perform EDA to understand the distribution of features, check for any patterns, and visualize the differences between churned and non-churned customers.

- **Model Building:**
- Use the labeled data to build a predictive model for churn. Consider models like logistic regression, decision trees, or more advanced techniques if needed.
- **Feature Importance:**
- Analyze the importance of different features in predicting churn to gain insights into what factors most influence customer retention.
- **Validation:**
- Validate your model using appropriate metrics (accuracy, precision, recall, F1 score) and perform cross-validation if necessary.
- **Documentation:**
- Document your findings and methodologies for future reference or reporting.

Summary of the steps and actions performed in the data cleaning process

1. **Churn Data Transformation**:
    - Converted churn values such that any value greater than 0 is mapped to 0 (inactive) and 0 is mapped to 1 (active).
    - Checked the distribution of the churn values, which showed a high imbalance with 94.24% inactive users and 5.76% active users.
2. **Column Exclusion**:
    - Dropped columns related to month 9 except for 'total_rech_data_9' and 'av_rech_amt_data_9'.
3. **Missing Values Analysis**:
    - Identified columns with more than 40% missing values and removed them.
    - Found that remaining missing values were less than 5%, so rows with missing values were dropped.
4. **Date Columns**:
    - Removed columns related to dates as they were deemed not useful for the analysis.
5. **Single Value Columns**:
    - Dropped columns with only a single unique value as they do not provide useful information for analysis.
6. **Mobile Number Column**:
    - Removed the 'mobile_number' column as it was not useful for the analysis.

- **Final Shape of Data**:
- The final dataset has 28,163 rows and 126 columns.

Summary of the tasks performed

1. **Remove Redundant Columns**:
   - **Code**: churn_data.drop('mobile_number', axis=1, inplace=True)
   - **Outcome**: You removed the mobile_number column from the dataframe, as it's likely an identifier not useful for analysis.
2. **Identify and Drop High Correlation Features**:
   - **Code**: Calculated the correlation matrix, identified high correlations, and dropped redundant columns.
   - **Outcome**: Removed columns with high correlation to reduce multicollinearity in the dataset. The resulting dataframe shape is (28163, 87).
3. **Generate Total MOU Features**:
   - **Code**: Created new columns total_mou_6, total_mou_7, and total_mou_8 by summing onnet_mou and offnet_mou columns for each month and then removed the original columns.
   - **Outcome**: Simplified the dataframe by combining onnet and offnet MOU into a single feature per month. The new dataframe shape is (28163, 84).
4. **Derive New Features for the Good Phase**:
   - **Code**: Created new features representing the "Good phase" by averaging the values from the 6th and 7th months and removed the original month-specific columns.
   - **Outcome**: Reduced redundancy by averaging certain features across two months, leading to a dataframe with fewer columns.
5. **Generate Additional Features**:
   - **Code**: Created a new feature gd_ph_vbc_3g from existing jul_vbc_3g and jun_vbc_3g columns and updated column names accordingly.
   - **Outcome**: Further reduced redundancy and consolidated similar features, resulting in a dataframe shape of (28163, 56).

Summary of the process

1. **Exploratory Data Analysis (EDA):**
   - **Churn by Tenure:** The scatter plot indicates that most churners have been with the service provider for less than 4 years.
   - **Correlation Heatmap:** Identified highly correlated variables, such as gd_phisd_og_mou & isd_og_mou_8, and gd_ph_arpu & arpu_8.
   - **Target Variable Distribution:** The distribution of the target variable churn shows significant imbalance, with 95% non-churners.
2. **Data Preparation:**
   - **Handling Outliers:** Applied upper limits to features with high outliers.
   - **Feature and Target Separation:** Separated features (X) from the target variable (y).
   - **Standardization:** Scaled the feature data using StandardScaler.
   - **Class Imbalance Handling:** Used SMOTE to balance the classes, resulting in equal numbers of churners and non-churners.
   - **PCA Transformation:** Applied PCA to reduce dimensionality and transform the features into 25 components.
3. **Model Building:**
   - **Logistic Regression:** Prepared data for logistic regression, utilizing the original features for RFE (Recursive Feature Elimination) to identify robust predictors of churn.
   - **Train-Test Split:** Split the data into training and testing sets (70% train, 30% test).

Model Evaluation:

- **Logistic Regression Results:** The regression results table shows coefficients for various features. For instance, arpu_8 has a positive coefficient, suggesting it is positively associated with churn.

Summary of the key steps and findings:

Variance Inflation Factor (VIF) Analysis:

- The VIF values for the features were calculated to assess multicollinearity.
- It was observed that gd_ph_total_mou had a very high VIF, indicating multicollinearity. Thus, it was decided to remove this feature from the model.

Model 2 (Revised GLM Model):

- After dropping gd_ph_total_mou, a new Generalized Linear Model (GLM) with a binomial family was constructed.
- The model summary shows that the updated GLM includes 24 features and has improved performance metrics.

Model 3 (Decision Tree):

- The model is giving accuracy level of 90%.

Model 4 (Random Forest):

- The model is giving train data accuracy of 91.5% and test data accuracy of 89.1%.

Conclusion

Considering our business problem of customer retention, prioritizing higher recall is crucial. Identifying potential churners accurately is more cost-effective than losing a customer and acquiring new ones.

Upon comparing the trained models, it is evident that the tuned Random Forrest perform exceptionally well, achieving accuracy of 91% & 89% on Train & Test data respectively. In this context, we choose the Random Forest model due to its simplicity and comparable performance.

Thank You