KENNESAW STATE UNIVERSITY

CPE 4040: Data Collection and Analysis, Spring 2023

Final Project: Predicting Customer Churn for a Telecommunication Company

Project Overview

You are a data analyst for a telecommunication company. The company wants to improve its customer churn rate by identifying factors that contribute to customers canceling their services. To achieve this, you have been provided a dataset of customer information, and have been tasked with building machine learning models to predict which customers are most likely to leave. Given the high cost of acquiring new customers compared to retaining existing ones, this exercise is critical for helping the company maintain its revenue and profitability.

Dataset

The dataset includes customer characteristics in the following categories:

- Demographic information: such as gender, age range, and if they have partners and dependents;
- Account information: such as tenure, contract type, and payment method.
- Services subscription: such as internet, phone, and TV, as well as their usage patterns and monthly charges.
- Churn: customers who left within the last month

You will find detailed description of the data columns in the Jupyter Notebook file.

Project Tasks

You will apply the skills and tools that you have learned throughout the semester to process, explore, and visualize the data. Additionally, you will also need to communicate your findings effectively, through charts and thoughtful comments about your analysis and conclusions.

To summarize, your analysis should include the following four sections.

1. **Data Cleaning and Preprocessing**:
   a. Examine the basic attributes and statistics of the dataset.
   b. Identify and handle missing values and outliers.
2. **Exploratory Data Analysis (EDA)**:
   a. Focus on examining the relationship between the features and the customer churn, and also analyze the distribution of certain features and the relationships between features.
   b. Use appropriate visualization techniques to explore the data.
   c. Present at least **five** charts and analyses.

3. **Predictive Modeling**:
    a. Perform feature engineering and selection to identify the relevant features for predicting customer churn.
    b. Follow step-by-step instructions to perform **logistic regression modeling** to classify if a customer is likely to churn.
    c. Evaluate the model performance using metrics such as accuracy, precision, recall, F1-score, etc.
    d. Interpret the coefficients and identify the most important features in the model.
4. **Conclusion**:
    a. Summarize the main findings and insights from the project.
    b. Suggest ways the company can improve the churn rate based on the analysis.

Deliverables

You will deliver a Jupyter Notebook file that includes your code, analysis, and comments. The Notebook also serves a report and, therefore, you will be required to use the Markdown feature in Jupyter to format the analysis to include headings, subheadings, and any other necessary formatting to make the report clear and concise.

A good reference: Markdown for Jupyter notebooks cheatsheet

Grading Rubric

Here is the grading rubric based on the four sections of the project.

- Data Cleaning and Preprocessing: 15 points
- Exploratory Data Analysis: 35 points
- Predictive Modeling: 35 points
- Conclusion: 10 points
- Overall Clarity and Organization: 5 points