Data Glacier

Data Science Intern at Data Glacier

Week 8: Deliverables

Name: Tejeswar Reddy Nalijeni

University: University of Cincinnati

Email: nalijety@mail.uc.edu or tejeswarreddyn2808@gmail.com

Country: United States

Specialization: Data Analyst

Batch Code: LISUM35

Date: Aug 26th, 2024

Submitted to: Data Glacier

Table of Contents:

1. Project Plan

Weeks	Date	Plan
Weeks 07	Aug 19, 2024	Project Preparation, Data Intake Report
Weeks 08	Aug 26, 2024	Data processing
Weeks 09	Sept 2, 2024	Data Processing (Advanced)
Weeks 10	Sept 9, 2024	Data Analysis, EDA
Weeks 11	Sept 16, 2024	Build Model Preparation
Weeks 12	Sept 23, 2024	Explore Different Model
Weeks 13	Sept 30, 2024	Presentation for data result & Model Evaluation, Code

2. Problem Statement and Understanding

2.1. Problem Description

The data is related to a company where they have various subscription plans based on tenure and provide support calls to customers if they have any issues or if there is any churn. Our goal is to predict the reasons for customer churn (variable y) and analyze the data to increase customer subscriptions.

2.2. Business Understanding

The main goal of this project is to predict the reasons behind customer churn in subscriptions using recorded data. This involves optimizing marketing efforts, improving customer engagement strategies, and ultimately boosting subscription numbers. By utilizing historical data for binary classification, the project seeks to accurately identify potential subscribers, understand the causes of churn, and increase customer retention through targeted and informed promotions.

3. Data Intake Report

Click here for **Data Intake Report**.

4. Data Understanding

4.1. Columns

1. Customer ID: A unique identifier for each customer.
2. Age: The age of the customer.
3. Gender: The gender of the customer.
4. Tenure: The length of time the customer has been with the service.
5. Usage Frequency: How often the customer uses the service.
6. Support Calls: The number of support calls made by the customer.
7. Payment Delay: The number of times the customer has delayed payment.

8. Subscription Type: The type of subscription the customer has.

9. Contract Length: The duration of the customer's contract.

10. Total Spend: The total amount of money the customer has spent.

11. Last Interaction: The time since the customer's last interaction with the service.

12. Churn: Indicates whether the customer has churned (likely a binary column with values like 0 for 'No' and 1 for 'Yes').

4.2. Information

As we haven't started the analysis yet, we can't determine if the data contains outliers, missing values, etc. However, we can outline a strategy for handling missing values and outliers.

Handling Missing Values:

1. If there are many missing values:
   - For categorical variables, replace missing values with the most common (mode) value.
   - For continuous variables, replace missing values with the median.
2. If the missing values are few:
   - Consider simply deleting those rows.

   So, as my data have only one missing row, I considered to delete that row which does not impact the result.

Handling Outliers:

This will be assessed on a case-by-case basis, depending on the dataset. We'll address this as we proceed with the analysis. “Age” appears normal but skewed to the left. “Total spend” appears normal bust skewed to right. Almost all the graphs except “Total spend` have strong Negative/left-skewed.

Campaign-Specific Data: The target variable for our predictive model is "y," indicating whether the customer churns or not.