

CRISPDM DOCUMENTATION

Author: George Nyandusi

PROJECT TITLE: SYRIATEL TELECOM COMPANY CHURN ANALYSIS AND PREDICTION

- CRISP DM(Cross-Industry Standard Process for Data Mining) refers to a popular methodology used in Data Mining, Data Analytics, and Data Science projects.
- It provides a detailed structure and iterative workflow for tackling Data Science projects.
- Some of the key stages involved in the CRISP DM framework include:
 - Business Understanding
 - Data Understanding
 - Data Preparation
 - Modeling
 - Evaluation
 - Deployment
- We will use the CRISP DM framework to analyze the Syriatel dataset in order to understand the reasons why there is customer churning.

Step 1: Business Understanding

Overview

SyriaTel, one of the biggest telecommunications company in Syria aims to be the best in the game by offering topnotch services to its customers. Although they are among the best, there is still a high customer turnover and this leads to loss of revenue for the company. This project aims at finding factors that influence churning and provide optimal recommendations that will ensure SyriaTel retains more customers and minimize losses.

Problem Statement

The company stakeholders wish to determine which features influence customer churning and derive insights that will enable them to make strategic decisions to curb the high churn rate.

Objectives

1. To develop a model that determines if a customer churns or not based on various attributes.
2. To identify the attributes that heavily impact if a customer is likely to churn or not.

Challenges

Some of the key challenges faced in analyzing the Syriatel telecom dataset include:

1. Data Imbalance – Churn datasets often have far fewer churned customers compared to retained ones, making it difficult for models to learn meaningful patterns.
2. Feature Selection – Identifying the most relevant features (e.g., call duration, international plans, customer service calls) is crucial, but telecom data can be noisy and high-dimensional.

Proposed Solution

This project aims to identify the features most associated with customer churning by analyzing data from Syriatel telecom dataset. The solution entails developing a comprehensive pipeline—from data gathering, cleaning, and Exploratory Data Analysis (EDA) to making models that help in extracting data-driven insights and recommendations that will support robust decision-making.

Conclusion

This analysis will empower the company with strategic recommendations on which features should be looked at to maximize the chances of retaining the highest number of customers and generating decent revenue.

Step 2: Data Understanding

- o The data for the project was obtained from the Kaggle website: <https://www.kaggle.com/datasets/becksddef/churn-in-telecoms-dataset>

- Some of the key attributes used in our analysis include:
 1. Area code
 2. International plan
 3. Voicemail plan
 4. Total day minutes

Data Checks

- Checked for missing values in the data.
- Checked for duplicates in our datasets.
- Checked the data types of each column in the datasets.
- Checked for uniformity of data in the datasets.

Step 3: Data Preparation

- In the data preparation phase, I aimed to clean and transform the data to obtain deeper insights that would aid in providing actionable recommendations.
- Some of the key data preprocessing steps I undertook include:
 - **Handling missing values:** I checked for missing values and duplicates and found none meaning the data was clean.

Step 4: Modeling

This project required me to make at least two models which I made using :

- a. Logistic Regression model
- b. Decision Tree classifier
- c. Random forest

Step 5: Evaluation

The evaluation metrics that were used include:

- a) Accuracy
- b) Precision
- c) Roc-Auc
- d) F1 score

Insights

- The main insights that were obtained from these tests are:

After my analysis I have found that the best model is the decision tree model.

the features that lead to higher churning have been found to include the total day charge, customer service calls, voicemail plan, international plan and total calls

Step 6: Deployment

- The analysis was carried out in a Jupyter Notebook.
- It was then saved and uploaded to the cloud (GitHub).