CRISPDM DOCUMENTATION

Author: Patrick Maina

PROJECT TITLE: ***CUSTOMER CHURN PREDICTION FOR SYRIATEL TELECOMMUNICATION COMPANY***

- CRISP DM (Cross-Industry Standard Process for Data Mining) refers to a popular methodology used in Data Analytics and Data Science Projects. It provides a detailed framework and iterative workflow for tackling Data Science Projects.

- Some of the fundamental stages involved in CRISP DM include:
    - Business Understanding
    - Data Understanding
    - Data Preparation
    - Modeling
    - Evaluation
    - Deployment

- We will use the CRISP DM framework to analyze the rate of customer churn in SyriaTel Telecommunication Company, and provide actionable insights and recommendations on how to minimize the customer churn rate.

Step 1: *Business Understanding*

Overview

One of the major challenges facing telecommunication providers is customer churn – a scenario where users discontinue their service, mostly due to dissatisfaction from the provider, or due to the availability of a better alternative from competitors. In order to mitigate this challenge, telecom companies are exploring churn prediction mechanisms, as well as the need to understand factors that contribute to customer churn.

Problem Statement

SyriaTel, a leading telecom provider, is experiencing a significant loss of customers who are choosing to leave its services for other competitors.

Objectives

1. To determine the key characteristics and behavior patterns that contribute to customer churn.

2. To develop a robust predictive model that will identify customers with a high likelihood of discontinuing their service, with a recall score of about 80%.

3. To provide data-driven insights and recommendations that will proactively engage, and retain high-risk customers.

Challenges

Some of the key challenges faced in analyzing customer churn rates from a telecommunication company include:

- Imbalance churn rates, which can make it difficult for predictive models to learn patterns related to churn, reducing the performance of the predictive models.

- The customer data may have data quality issues such as missing, inconsistent or outdated information, especially when sourced from multiple systems such as billing, customer support, or usage logs. This significantly impacts the reliability of the analysis.

- It is difficult to capture and model accurately churn behavior, due to its impact by both observable and unobservable factors, such as customer satisfaction, competitive offers, or personal circumstances.

Proposed Solution

This project aims to analyze the customer churn dataset, identify the key features that contribute to customer churn, and build a predictive model that is able to predict the likelihood of a customer churning from the company. The solution entails a number of crucial steps—from data exploration, data manipulation, Exploratory Data Analysis (EDA), data preprocessing, modeling, model evaluation, model tuning, and extracting data-driven insights and recommendations that will support robust decision-making.

Conclusion

This analysis and modeling will empower the SyriaTel Telecommunication Company with strategic recommendations on how to mitigate customer churn, promote long-term customer retention and enhance long-term profitability.

Step 2: *Data Understanding*

- The Churn in Telecom's dataset was collected from Kaggle. It contains about ***3333 records*****,** and ***21 columns*****,** of which ***4 columns*** are categorical columns, while the remaining ***17 columns*** are numerical columns.

- It contains essential customer churn attributes such as:
    - **State and Area Code:** Contains different states and area codes of the customers subscribed to SyriaTel, both who are churning and who are not churning from the company.
    - **International and Voice Mail Plans:** It gives a directive as to which customers are subscribed to either an International plan, a voice mail plan, both, or none.
    - **Call rates:** It provides information on the different customer call rates for day, evening, night, and international calls.

    - **Customer Service calls:** It provides information on the number of customer service calls the customers are experiencing with the support staff in the company.

- **Data Checks:**
    - Checked for missing values in the dataset
    - Checked for duplicate rows and inconsistency in the dataset.
    - Checked the data type of each column in the dataset.
    - Checked for uniformity of data in the dataset.

Step 3: *Data Preparation*

- In the data preparation phase, we aimed to clean the data, perform EDA to understand the relationship between the features and the target, and to preprocess the data for preparation for the modeling phase.

- Some of the key data preparation steps taken were:
    - **Handling missing values:** The dataset did not contain any missing values or duplicate rows, but this check had to be performed for due diligence.
    - **Standardizing column names:** All the columns were uniformly standardized using the following methods:
        - **strip():** removes trailing whitespaces in a string
        - **lower():** converts a string to lowercase
        - **replace():** replaces characters in a string
    - **Plotting the distributions:** we performed EDA by plotting some of the key features, and visualizing the relationships between these features and the target variable. Some of the crucial visualizations we did include:
        - **Customer churn distribution:** This plot illustrates the rate of customer churn from the company. From the plot, out of the 3333 customers in the dataset, only ***483 customers*** supposedly churned from the company, which is about ***14.5%***. This shows a huge imbalance in the churn and not churn classes.

        - **Distribution of customers by Area Code:** From this plot, area code ***415*** had the highest number of customers with about ***1655 customers***, which accounts for about ***49.7%.*** Area codes ***408*** and ***510*** had about ***838 customers*** and ***840 customers*** respectively.
        - **Relationship between customer service calls and the churn rate, by area code:** From the plot**,** customers who churn from the company tend to have more customer service calls than those who don't. Additionally, the majority of the customers who churn come from area codes ***415*** and ***510.***
    - **Data Preprocessing:** Some data preprocessing steps were taken to prepare the data for modeling. The key preprocessing steps that were done include:
        - **One-Hot Encoding:** This encoding scheme was performed on the categorical columns (State, Area code, International Plan, and Voice Mail plan) to create new columns for each category where **1** implies that the category is present, and **0** implies that the category is absent in the specific columns.
        - **Label Encoding:** This encoding scheme was performed on the target variable (Churn), where the categorical values (True/False) were converted into numerical labels (1/0) respectively.
        - **Feature Scaling:** The numerical features were scaled to a range of (0,1) to ensure all the columns have a standard range of values, a necessary step taken in modeling.

Step 4: *Modeling*

- In the modeling phase, we performed several key steps to ensure the modeling process was successful.

- These steps include:
    - Defined the X(features) and y(target) variables.
    - Resampled the data to handle class imbalance in the target variable using **SMOTE (Synthesis Over-Sampling Technique)**

    - Implemented the **train_test_split()** method for splitting the dataset into training and testing sets (80/20 split)
    - Trained six different Machine Learning models. These models include:
        - Logistic Regression
        - Decision Tree
        - Random Forest
        - XGBoost
        - K-Nearest Neighbor
        - Gradient Boosting
    - Generated a classification report that contained important performance metrics such as **recall** and **model accuracy.**
    - Plotted a **confusion matrix** for all the models to evaluate the rate of true and false positives and negatives.

Step 5: *Evaluation*

- In the evaluation phase, we used the **recall** and **ROC_AUC score** to measure the performance of the six models.

- This was done by:
    - Plotting the **ROC (Receiver Operation Characteristic)** curves and computing the **AUC (Area Under the Curve)** scores for all the six models, and doing a comparison of the curves and the AUC scores.
    - Computing the **recall** score for all the six models, and comparing the scores of each model.

- From the evaluation, the top 3 models based on recall score were:
    - **Gradient Boosting - 0.807**
    - **XGBoost - 0.795**
    - **Decision Tree - 0.75**

- And the top 3 models based on ROC_AUC score were:
    - **Gradient Boosting - 0.912**
    - **XGBoost - 0.910**

    - **Random Forest - 0.908**

- Based on these results, we tuned the Gradient Boosting and XGBoost models to enhance performance. The results of the tuned models were as follows:
    - The tuned Gradient Boosting model achieved a recall score of **0.81**, which was similar to the untuned model recall score, and an AUC score of **0.921**, which was an improvement from the untuned model AUC score.
    - The tuned XGBoost model achieved a recall score of **0.82**, which was a significant improvement from the untuned model recall score, and an AUC score of **0.911**, which was a slight improvement from the untuned model AUC score.

- We can therefore conclude that the **XGBoost model** was the best performing model, based on the recall score, which was the main metric for this prediction modeling task.

- Some of the data-driven recommendations and insights obtained from this analysis include:
    - Offering specialized discounts, loyalty rewards, or exclusive promotions in high-churn area codes such as **415**, which will serve as an effective incentive to retain customers.
    - Investing in comprehensive training programs for support staff, and implementing better issue/conflict resolution frameworks in order to enhance customer satisfaction, thus minimizing the rate of churn.
    - Developing localized market efforts, personalized engagement strategies, and enhanced customer support in high-churn areas such as Texas and New Jersey, in order to strengthen customer loyalty and retention.

Step 6: *Deployment.*

- The analysis and modeling were carried out in a Jupyter Notebook.

- The notebook was then saved and uploaded to a cloud repository (GitHub).