DSFT-04 - Full Time - Phase 03
STELLA CHEROTICH K.
GITHUB REPO // Medium Blog

CUSTOMER CHURN PREDICTIVE MODEL DOCUMENTATION

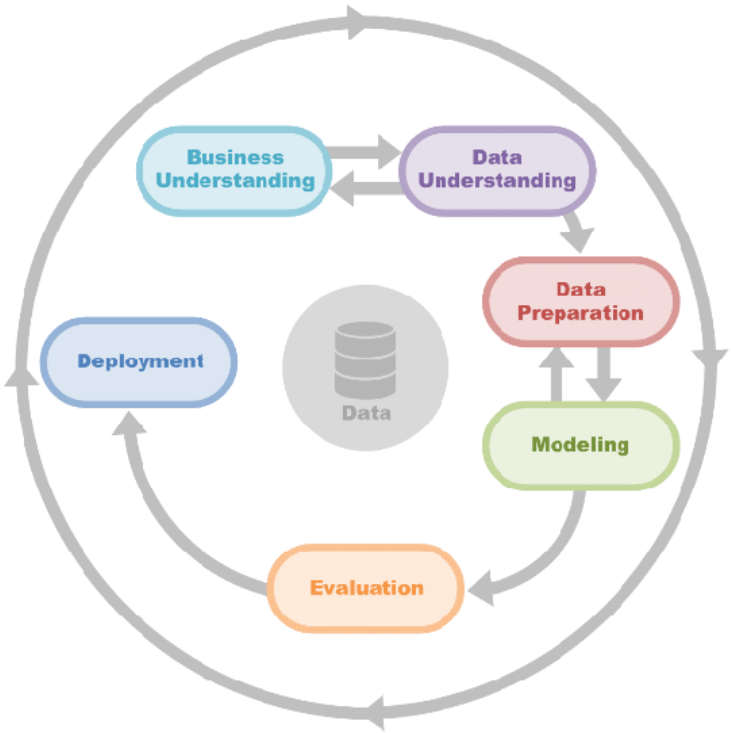
Project Overview

The objective of this project was to develop a binary classification model to predict whether a customer of SyriaTel, a telecommunications company, is likely to stop doing business in the near future. The primary goal was to identify predictable patterns in customer behavior in order to help the company reduce financial losses associated with customer churn.
Stakeholder: SyriaTel

Business Understanding

Customer churn has emerged as a critical concern for companies like SyriaTel in the fiercely competitive telecommunications industry. With customers having numerous options and increasing expectations, retaining existing customers has become paramount.
Churn not only leads to immediate **revenue loss** but also exerts significant pressure on customer acquisition costs. Understanding the factors that contribute to churn and being able to predict it with accuracy is crucial for telecom businesses to develop effective retention strategies. By analyzing historical customer data, telecom companies can gain valuable insights into customer behavior, preferences, and interactions, enabling them to **identify** potential churners and tailor retention efforts accordingly and proactively.
This proactive approach minimizes revenue loss and enhances customer satisfaction, loyalty, and overall business performance.

The Data Science Process that the analysis follows is the CRISP-DM process.



Problem Statement

SyriaTel would like to maintain/increase the customer retention rate as well as seeking to address the challenge of customer churn by developing an accurate binary classification model that predicts the likelihood of customers discontinuing their services.

Objectives

The objective of this analysis is to:

1. Develop a highly accurate binary classification model that predicts customer churn for SyriaTel.
2. Identify predictable patterns and insights in customer behavior to proactively identify customers at a high risk of churning.
3. Enable SyriaTel to optimize retention strategies, allocate resources effectively, and minimize financial losses associated with customer churn.

Data Understanding

The SyriaTel Dataset was retrieved from Kaggle and can be found here.
The original dataset contains 3333 rows and 21 columns. The columns included information that is associated with features of the customer information such as:

1. state
2. account length
3. area code
4. phone number
5. international plan
6. voice mail plan
7. number vmail messages
8. total day minutes
9. total day calls
10. total day charge
11. total eve minutes
12. total eve calls
13. total eve charge
14. total night minutes
15. total night calls
16. total night charge
17. total intl minutes
18. total intl calls
19. total intl charge
20. customer service calls
21. churn
22. Total Expenditure[1]

[1] This column was added as a new column by the author through feature engineering

To understand the data further, EDA (Exploratory Data Analysis) was conducted especially looking at the **target variable** which was churn, and how other variables influenced it. Additionally, a correlation matrix was made in order to see what features were highly correlated.

Data Preparation

As the data had no missing values or any duplicate values, the data instead was cleaned: this included dropping certain columns from the dataset, transforming the data, and preprocessing it so that it'd be suitable for the purpose of running the various models. Other methods that were used in this stage included:

- Normalizing the dataset
- Data type conversions
- Dealing with multicollinearity
- Setting the target variables and splitting the train and test data

Modeling

In this stage, statistical and machine learning models are developed using the preprocessed data. Four (4) models were created:

1. Logistic regression model – Baseline model
2. Decision Tree Model.
3. Random Forest Model
4. Support Vector Machine

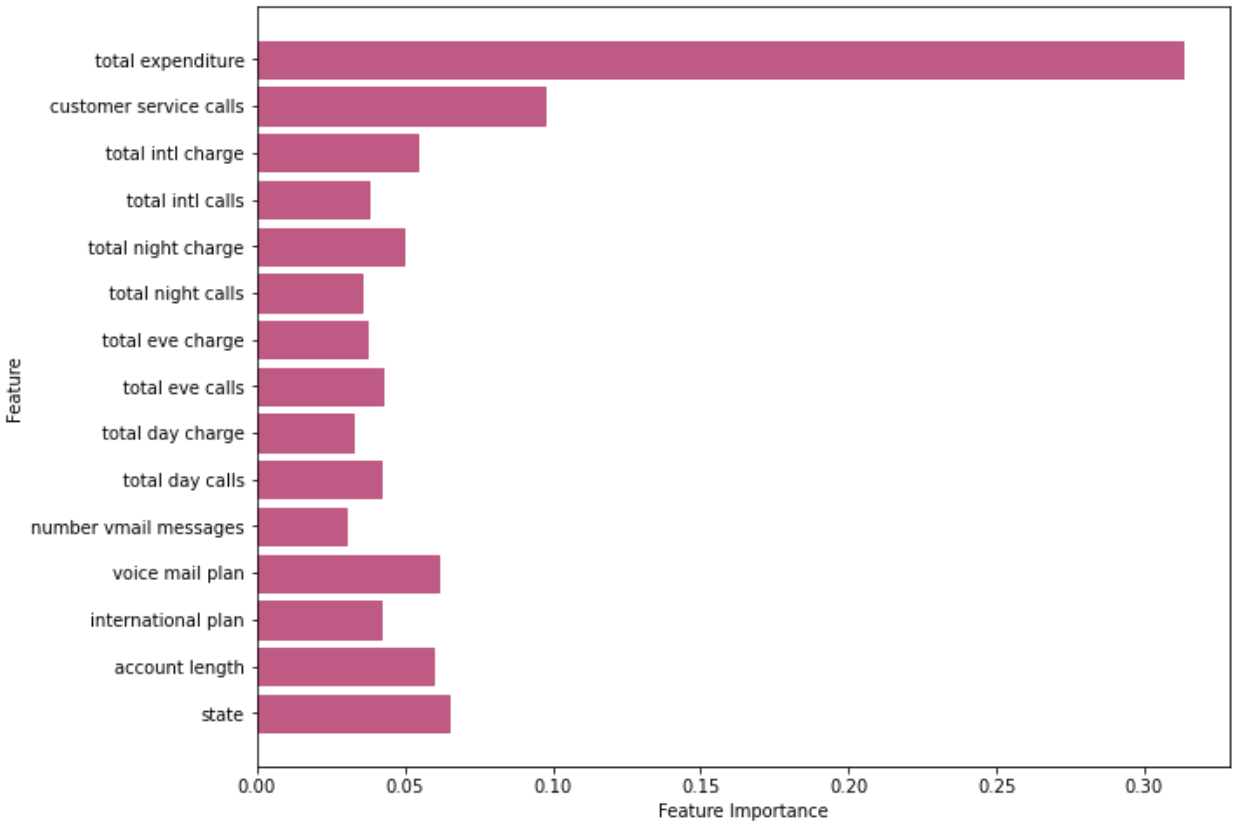
Evaluation

The best model out of the four based on the Recall Score was the Decision Tree Model as it had a percentage value of 0.8649; the scores for all the models are illustrated in the table below:

Model	Accuracy	Precision	F1 Score	Recall
Logistic Regression	0.6978	0.2689	0.3942	0.7387
Decision Tree	0.8465	0.4593	0.6000	**0.8649**
Random Forest	0.9472	0.8384	0.7905	0.7477
SVM	0.8717	0.5141	0.5771	0.6577

Additionally, the key features that were shown to influence whether a customer would churn or not, that is displayed below. We can note that the total expenditure is a key predicting variable as to whether or not a customer will churn, as well as the customer service calls made and the services provided in certain plans, such as voice mail plan and international plans.

The recommndations will be outlined in the next and final section



Recommendations

Based on the model results, as the Data Scientist assigned to this project, I would recommend the following.

1. As **total expenditure** is an influencing factor for whether or not a customer will churn; It is important that SyriaTel reconsiders some of the service costs, perhaps in a way that would be more accomodating to individuals that have a certain budget.
2. Additionally, focus should be placed on the issues that are raised during the **customer service calls** , while also ensuring that those who are responding to the customers needs are adequately trained as well as adhering to good customer service norms, in order to ensure quality service is provided.
3. Furthermore, SyriaTel should consider taking a customer-centered approach, for example having certain plans that can be modified to suit the needs of the diverse customer base, example: some customers may be more interested in the international plan compared to having a voice mail plan.