Phase 3 Project Report

Title: SyriaTel Telecommunications Company Churn Prediction.

Project Overview

SyriaTel is a telecommunications company that prides itself in offering top-notch services to their customers. They are the leading telecoms company in their country and want to remain the leader in that particular sphere. Over the years they have chartered a lot of the strides in technology in their country and want to continue improving.

1.Business Understanding

Stakeholders: SyriaTel executives and managers

Problem Statement

In a bid to continue leading, SyriaTel was facing a significant challenge, CUSTOMER CHURN i.e. where the customers were discontinuing their services and switching to other service providers. This churn not only lead to revenue loss but also affected the company's market position and customer satisfaction.

Proposed Solution

The proposed solution was to develop a machine learning model that can analyse customer data, including demographics, usage patterns, service subscriptions, to predict the likelihood of customer churn. The model was to be able to identify customers who are most likely to churn, enabling the telecommunication company to take proactive measures to retain them.

Project Objectives

1. To develop a model that will help in predicting 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.

Project scope and limitations

- This project was orchestrated as an extra advisory tool to support top-level management make informed decisions to deal with customer retention.
- The project outputs i.e. The model was not be realized as a full application with a user interface but rather a final report on the findings based on the data used which include a number of recommendations.
- Internal data from the company was the primary data source that drove this project.
- Ultimately the final steps taken to mitigate the situation was up to the company.

Benchmark metric

- The bench mark evaluation metric that will be used in this project is **ACCURACY**.
- **Justification**: from objective 1 we want to know if a customer churns therefore accuracy and F1 score would be suitable.
- F1-score is also considered but not the main metric.

2.Data Understanding

Here we explored the data to get a better understanding of its state, then decided on the steps needed to take to clean it. We begun by defining a class for the following tasks:

- getting the shape of the data
- getting data info
- descriptive stats

From the class describer, the dataset had:

- 3333 customers.
- 21 customer features: 4 string predictors, 16 numeric predictors and the target.
- Various transformations were applied on the dataset both for analysis and modelling e.g. type conversions, feature selection etc.

3.Data preparation

In this section we cleaned the dataset by dealing with:

- Missing values
- Duplicated values
- Outliers
- Inconsistencies in the dataset

We created a class to handle the cleaning process. The class was able to identify missing values, duplicates both generally and using a unique column.

- The Data had no missing values.
- The Data had no duplicates.
- The unique column Phone Number had no duplicates
- Outliers were not removed.
- They were true outliers. Most of the outliers revolved around talk duration on the phone. It is plausible that a customer may talk for 350 minutes in a day either continuously or discretely.
- The length of the data was too small to drop the outliers ergo were not touched.

4. Exploratory Data Analysis

In this section, we explored univariate EDA and bivariate EDA.

- Overall churn ratio.
- Churn against states.
- Churn against international plan.
- Churn against voice mail plan.
- Churn against number of customer service calls.

From the analysis above it was evident that the areas explored need to be considered when improving services of the company. More features were explored in modelling to add on to the recommendations.

5. Pre-processing

In this section we prepared the data for modelling.

- Some of the pre-processing that took place here include:
  - Feature selection.
  - Train test split.
  - Encoding: dummy encoding and basic replace application.
- We investigated the important features of the data set and chose them to reduce complexity of the models ergo avoiding overfitting from the get go.
- The criteria used to select relevant columns is **domain knowledge and feature importance analysis provided by decision trees.**

6. Modelling

- In this section is where the magic happened.
- The problem at hand was a classification problem.
- We explored 3 models: a baseline DecisionTreeClassifier, a randomforest classifier and a tuned random forest model.
- Model accuracy and F1 score will be the metrics for evaluation.

  Justification: Accuracy to get a verdict if a customer churn's or not. F1 score to get a balance between precision and recall.

- Accuracy of 70% and F1 score of 70% will be the threshold to deem the model as successful.

Machine Learning Communication - baseline model

Rationale why modelling was implemented.

- While simpler forms of data analysis, such as descriptive statistics or basic data visualization, provided initial insights, they were not sufficient for complex problems or dataset. Machine learning leveraged advanced algorithms to uncover hidden patterns.

Results.

- Accuracy on the training set: 1.0
- F1 score of the model on train set: 1.0

- Accuracy on the testing set: 0.93
- F1 score of the model on test set: 0.76

- The model was **overfitting**

- The accuracy meant that the model could predict with an accuracy of **93%** whether a customer will churn or not.

Limitations of baseline model

- The current model was not fit for prediction since it was not generalizing well to new data even with high accuracy. The model was **overfitting**.
- This we saw from the 7% difference between train and test accuracy.
- This we saw from the large difference between train and test F1-score

Model 2

- Built an ensemble model with hyperparameters.
- Different from base model in that we employed an ensemble model with hyperparameters to battle overfitting, employing SMOTE to handle the imbalance issue and scaling down of features also to battle overfitting.

Rationale why ensemble modelling was implemented.

- While simpler forms of data analysis, such as descriptive statistics or basic data visualization or baseline model could provide initial insights, they were not sufficient for the problem or dataset such as this one. Ensemble models leveraged advanced algorithms to uncover hidden patterns, make accurate predictions, and provide actionable insights that greatly benefitted SyriaTel in decision-making processes.

Machine Learning Communication for Model 2.

Results.

- Accuracy on the training set: 0.92
- F1 score of the model on train set: 0.75

- Accuracy on the testing set: 0.91
- F1 score of the model on test set: 0.72

- The model was **not overfitting** but it is exhibiting issues because the test scores are higher than train. This may be due to data leakage, random variations or training size.

- The accuracy meant that the model predicted with an accuracy of **91%** whether a customer would churn or not.

Limitations.

- The current model was not fit for prediction since it is experiencing higher test scores indicating internal issues e.g. random variations, data leakage etc.
- This we saw from the difference between train and test accuracy.
- This we saw from the difference between train and test F1-score.

Model 3

- Built a model using gridsearchCV to find best set of hyper-parameters.
- Different from base model and Model 2 in that we employed an ensemble model with gridsearchCV to find the best set of hyperparameters to battle overfitting while improving accuracy and F1 score.
- Employed SMOTE to handle the imbalance issue and scaling down of features also to battle overfitting.
- Employed cross validation to tackle the random variation issue.

Rationale why tuned ensemble modelling was implemented using gridsearchCV.

- While simpler forms of modelling and data visualization provided initial insights, they were not sufficient for battling overfitting. Grid search and tuned Ensemble models leveraged advanced algorithms to uncover hidden patterns, make accurate predictions, provide actionable insights and battle overfitting by finding the best set of hyperparameters. This greatly benefitted SyriaTel in decision-making processes.

Machine Learning Communication for Model 3.

Results.

- Accuracy on the training set: 0.93
- F1 score of the model on train set: 0.78

- Accuracy on the testing set: 0.92
- F1 score of the model on test set: 0.74

- The model was **not overfitting**. It was now performing as expected.

- The accuracy meant that the model can predict with an accuracy of **92%** whether a customer would churn or not.
  .

Best Model: Model 3

Justification.

- Model 3 proved to have a balanced performance on train and test instances. It demonstrated proper metrics with accuracy score of 92%.
- Overall it did not overfit.

7. Findings and Recommendations

7.1. Findings.

The following findings were found:

1. From the modelling exercise, the 3rd model was the best and was able to predict the target with an accuracy of 92%. This means given customer features, it is able to predict churn/no churn with an accuracy of 92%.
2. The features that heavily determine churn are **total day minutes, total day charge, customer service calls**
3. Findings from EDA are also considered.

7.2. Recommendations.

The following recommendations were made based on the whole exercise:

1. Improve Service Quality and Customer Experience: High customer service calls may indicate that customers are experiencing issues or dissatisfaction with the service. To reduce churn, the company should focus on improving service quality, addressing customer concerns promptly, and providing excellent customer support. This can be achieved through staff training, efficient complaint resolution processes, and regular feedback collection to identify and rectify service gaps.
2. Review Pricing Strategies: Since total day minutes and total day charge are influential factors in churn, it is crucial to evaluate the pricing structure and competitiveness. Consider conducting market research and competitor analysis to ensure that the company's pricing is competitive and aligned with customers' expectations. Offering attractive plans, discounts, or incentives for loyal customers can help retain them and discourage churn.
3. Proactive Customer Engagement and Retention Programs: Rather than waiting for customers to reach out with issues or complaints, the company can proactively engage customers through personalized communication and retention programs. This can include sending targeted offers, exclusive promotions, and customized recommendations based on customers' usage patterns and preferences. Building strong relationships with customers and providing them with incentives to stay can significantly reduce churn rates.
4. Analyse Churn Patterns and Predictive Modelling: Add more data on churn to analyse patterns and trends. Implement predictive modelling techniques in the current systems, to forecast customer churn probability based on various features. By identifying customers who are at high risk of churn, the company can proactively reach out to them with targeted retention strategies and offers, increasing the chances of retaining those customers.

Next steps

- 1. Implement the recommendations stated.
- 2. Gather more data for modeling to improve accuracy.