DATA REPORT

CRISP-DM METHODOLOGY

CUSTOMER CHURN PREDICTION FOR SYRIATEL TELECOMMUNICATION COMPANY

GROUP 5 MEMBERS

MUSI CALORI

JESSICA GICHIMU

IVY VICKER

BOB LEWIS

SEPTEMBER 2025

TABLE OF CONTENTS

1. Business Understanding

1.1 Business Overview

SyriaTel is a telecommunication company whose revenue model depends on recurring subscriptions. In the competitive industry, customer churn is a major threat as it reduces revenue, raises acquisition costs and lowers customer lifetime value. Minimizing churn is therefore important for sustaining profitability and customer loyalty.

1.2 Problem Statement

SyriaTel is experiencing a high rate of customer churn, leading to significant revenue loss and reduced competitiveness. Without clear insight into which customers are most at risk and the factors driving their decisions, the company struggles to implement effective retention strategies.

1.3 Business Objectives

1.3.1 Main Objective

The main objective of this project is to develop a classification model that predicts customer churn for SyriaTel and identifies the key factors influencing it. The results will help reduce churn, improve customer retention and protect long term profitability.

1.3.2 Specific Objectives

To achieve the main objective, the project has the following specific objectives:

1. Develop a churn prediction model and evaluate its performance using relevant metrics.
2. Identify the key features and customer characteristics that significantly influence churn.
3. Compare different classification models to select the one that best balances predictive performance with business needs.

These objectives guided the modeling and evaluation stages of the CRISP-DM process.

1.3.3 Research Questions

To ensure the analysis directly addresses SyriaTel's business problem, the following research questions were defined:

1. Can a churn prediction model achieve strong performance?
2. Which features and customer characteristics have the greatest influence on churn?
3. Which classification model provides the best balance between predictive performance and business applicability?

The project answers these research questions through data preparation, model development, evaluation and interpretation to provide data driven insights that support customer retention strategies.

1.4 Success Criteria

Project success will be measured by both model performance and business outcomes. The model should accurately identify customers at risk of churning so that SyriaTel can act in time. From a business perspective, success means achieving a measurable reduction in churn compared to current levels, improving retention and customer lifetime value, and strengthening overall profitability.

2. Data Understanding

2.1 Data Source

The dataset used in this project contains customer level information provided by SyriaTel. It includes 3,333 rows with 21 columns describing customer demographics, service usage, billing and interactions with customer service. The target variable shows whether a customer churned.

2.2 Data Description

The dataset contains both categorical and numerical variables. Key features include:

- **Categorical variables:** State, Area Code, International Plan and Voice Mail Plan.
- **Numerical variables:** Account Length, Number of Voice Mail Messages, Customer Service Calls, and Call counts and Charges. In addition, Day, Eve, Night and International Minutes.
- **Target variable:** Churn with binary of Yes/No.

The distribution of the target shows that 85.5% of customers did not churn while 14.5% did churn. This shows a class imbalance that must be addressed during modeling.

2.3 Data Quality Checks

The dataset was assessed for quality before modeling.

- **Missing values:** None were detected across all 21 variables.
- **Duplicates:** No duplicate records were found.
- **Consistency:** Variable names were standardized such as Total_Day_Minutes and Customer_Service_Calls to ensure clarity.

- **Correlations:** Strong correlations were identified between usage minutes and their corresponding charges such as Day Minutes and Day Charge. Highly correlated features were flagged for removal to avoid redundancy.
- **Class imbalance:** The target variable is imbalanced with churners making up only 14.5% of the dataset. This imbalance requires careful handling during modeling.

3. Data Preparation

3.1 Data Cleaning

To prepare the data for modeling the following steps were taken:

- Highly correlated features such as usage minutes and their corresponding charges were dropped to avoid redundancy.
- The target variable was confirmed as binary (Yes/No) and categorical variables were standardized for consistency.

These steps ensured that the dataset was ready for transformation and modeling.

3.2 Feature Engineering and Transformation

Categorical variables such as State, International Plan and Voice Mail Plan were converted into numerical form using one-hot encoding to make them suitable for modeling.

In addition, numerical variables were retained in their original scales since the chosen models are not sensitive to feature scaling. The dataset was then split into training and testing sets to enable model development and evaluation.

These transformations ensured that both categorical and numerical information could be effectively used during modeling.

3.3 Handling Class Imbalance

The dataset showed a high imbalance with churners representing only 14.5% of the customers. To address this, class weights were applied during model training so that misclassified churners carried a higher penalty than non churners. This approach helped the models focus on correctly identifying at risk customers which improved recall on the churn class.

4. Modeling

4.1 Overview of Models

Several classification models were tested to predict customer churn. These included:

- Logistic Regression as a baseline model for interpretability.
- Random Forest to show nonlinear relationships and feature interactions.
- Gradient Boosting Classifier to improve performance by combining multiple weak learners sequentially.

These models were selected to balance interpretability, predictive performance and practical application for SyriaTel's business needs.

4.2 Model Training

The dataset was split into training and testing sets to enable fair evaluation. Each model was trained on the training set using cross validation to ensure robustness. Hyperparameters were tuned to improve performance with a focus on recall for the churn class due to the business importance of identifying at risk customers. Class weights were applied during training to address the imbalance in the target variable. This ensured that models were optimized to prioritize churn detection.

4.3 Model Assessment

Model performance was evaluated on the test set using accuracy, precision, recall, F1-score and ROC-AUC. Logistic Regression served as the baseline, performing reasonably but with limited ability to show complex patterns. Random Forest improved overall accuracy and provided valuable feature importance insights. The Gradient Boosting Classifier delivered the strongest balance of performance. It achieved high recall on the churn class while maintaining a competitive F1-score and overall accuracy.

Based on these results, Gradient Boosting was selected as the best performing model. This is because it aligned most closely with the project's success criteria of correctly identifying churners while supporting actionable business insights. This model provides SyriaTel with a reliable tool for guiding customer retention strategies.

5. Evaluation

5.1 Evaluate Results

The Gradient Boosting Classifier outperformed the baseline Logistic Regression and Random Forest models. It achieved strong recall on the churn class ensuring that most at risk customers were identified, while maintaining balanced precision, F1-score and overall accuracy. These results show that the model meets the technical success criteria set earlier in the project.

5.2 Review Process

The project objectives and research questions were addressed successfully. A predictive model was developed, evaluated and selected based on its ability to balance recall with overall performance. Key drivers of churn were identified including customer service calls, international plan and high day time usage charges. No major gaps were identified in the process confirming that the methodology was applied consistently.

5.3 Determine Next Steps

Based on the results, the project is ready to proceed to the Deployment phase. SyriaTel can use the model to identify high risk customers and direct retention resources more effectively. In addition, the insights on churn drivers suggest practical interventions. These could include addressing frequent service issues and reviewing plan pricing to further reduce churn.

6. Deployment

6.1 Plan Deployment

The selected Gradient Boosting model can be deployed to support SyriaTel's customer retention strategy. It will be used to generate churn risk scores that allow the business to identify customers most likely to leave. These scores can then be added into existing customer relationship management systems to guide targeted retention campaigns.

6.2 Monitoring and Maintenance

Model performance should be tracked over time to ensure reliability. Regular monitoring will confirm that recall on the churn class remains strong and that predictions remain consistent as customer behavior or market conditions change. Periodic retraining with updated data will be necessary to maintain accuracy.

6.3 Final Report

This data report provides SyriaTel with a clear overview of the project's findings, methodology and recommendations. It should be shared with the stakeholder to support decision making.

6.4 Review Project

The project applied the CRISP-DM methodology to SyriaTel's churn problem. The process delivered a high performing predictive model, identified key churn drivers and produced actionable recommendations. The key insights include the importance of addressing class imbalance early, prioritizing recall and aligning modeling with business objectives. For reproducibility, the Jupyter notebook and the data report were uploaded to GitHub, a cloud repository.

7. Conclusion and Recommendations

7.1 Conclusion

This project applied the CRISP-DM methodology to predict customer churn for SyriaTel. Gradient Boosting emerged as the best performing model as it achieved strong recall on the churn class while maintaining balanced overall performance. The project met its objectives by building a robust predictive model, identifying key churn drivers and providing insights that support SyriaTel's business goal of reducing customer churn.

7.2 Recommendations

1. Target high risk customers by using the model's predictions to prioritize interventions for customers most likely to churn.
2. Improve service quality by addressing frequent customer service issues. This is due to high service call volumes being strongly linked to churn.
3. Reassess plan offerings by reviewing international plan pricing and structure. This is because these customers showed higher churn rates.
4. Maintain and update the model by continuously monitoring performance. In addition, retraining the model with new data to ensure long term reliability.

These actionable recommendations will help support informed, evidence based decision-making as SyriaTel works to reduce churn, improve retention and strengthen long term customer loyalty and profitability.