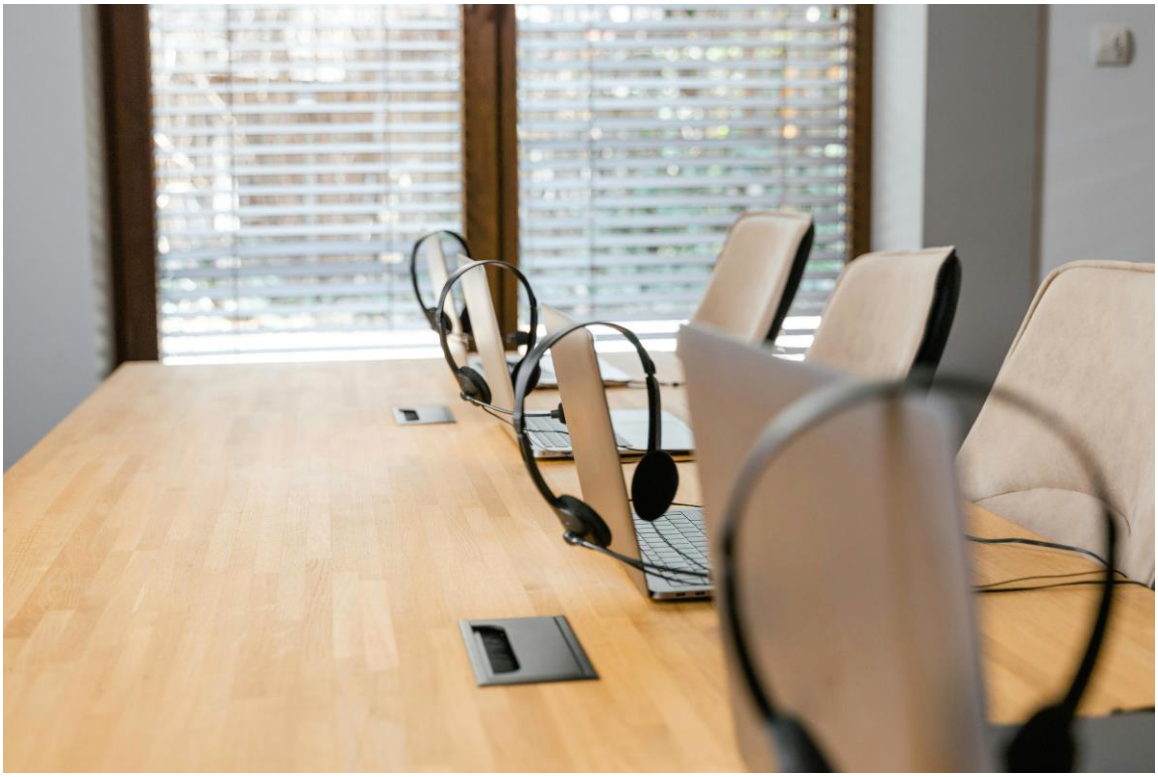
PREDICTING CUSTOMER CHURN FOR ENHANCED RETENTION STRATEGIES AT SYRIATEL TELECOMUNICATION



By Michelle Anyango

1: BUSINESS UNDERSTANDING

Business Problem

Telecommunication companies need to attract new customers and retain existing ones to grow their revenue. Customer churn, where customers cancel their service, is a major concern. This can be caused by factors like better pricing, poor service, or lack of engagement. I understand that keeping current customers is more cost-effective than constantly acquiring new ones. This project aimed to create a model that predicts customer churn and identifies key factors, helping SyriaTel take action to reduce churn.

Problem Statement

The goal of this project was to develop a model that predicts whether a customer is likely to churn from SyriaTel, a telecommunications company. This binary classification task analyzed customer behavior and demographic data to identify patterns that may signal a higher risk of churn. The objective was to help SyriaTel reduce the financial impact of churn by implementing proactive retention strategies. Additionally, I aimed to create a reliable churn prediction model by thoroughly analyzing key features based on historical data from the company.

2: DATA UNDERSTANDING

2.1 Key Details

I used the Churn in Telecoms Dataset from Kaggle, which contains data on customer behavior and their likelihood of churning in a telecom company.

Rows: 3,333 (each row represents a customer)

Columns: 21 (each column represents a customer feature)

2.2 Dataset Features

Customer Information: Includes details like customer state and phone number.

Account Information: Includes subscription plan and service features.

Usage Metrics: Covers phone usage details like minutes and charges.

Customer Service Interaction: Tracks the number of customer service calls.

Target Variable: Indicates whether the customer churned (Churn).

3: DATA PREPARATION

3.1 Libraries

I imported the required libraries for data analysis and modeling. These libraries include tools for data manipulation (pandas, numpy), visualization (matplotlib, seaborn), machine learning (scikit-learn), and handling imbalanced data (SMOTE).

3.2 Loading the Data and Understanding

I loaded the dataset from the "Churn_In_Telecom.csv" file for analysis. I also checked the shape and information about the dataset.

3.3 Data Preparation

Checked for missing values and identified any duplicate phone numbers in the dataset.

Replaced False/True with 0, 1 in the churn column to prepare the data for modeling.

Filtered the dataset to include only numerical columns for analysis.

3.4 Visualizing Customer Plans Distribution

I created two pie charts to visualize the distribution of customer subscriptions for different plans. The first pie chart shows the distribution of customers who have subscribed to the International Plan (Yes or No), and the second shows the distribution of customers who have subscribed to the VoiceMail Plan (Yes or No).

3.5 Visualizing Total Minutes Distribution

I summed up the total minutes spent by customers during different times of the day for the following categories: Total Day Minutes, Total Evening Minutes, Total Night Minutes, and Total International Minutes. The total minutes for each category were then visualized using a bar chart.

3.6 Visualizing Churn Distribution

I created a pie chart to show the distribution of customer churn, which reveals that a smaller percentage of customers are churning compared to those who remain with the company.

3.7 Grouping Data by State and Churn

I grouped the customer data by state and churn status, mapped the state initials to their full names, and calculated the total number of customers in each state.

3.8 Churned Customers by State

I focused on customers who churned and visualized the number of churned customers in each state. This helped identify regions with higher churn rates.

3.9 Customers Who Remained by State

I focused on customers who remained with the service and visualized the number of customers who stayed in each state.

3.10 Correlation Matrix Analysis

I created a correlation matrix to understand relationships between numerical features. This helped identify strong correlations, guiding feature selection.

3.11 Average Number of International Calls by Churn Status

I calculated the average number of international calls for customers who churned and those who didn't to see if this has an impact on churn.

3.12 Churn Rate Percentage by Customer Service Calls

I calculated the churn rate percentage for each number of customer service calls, visualizing how customer service interactions relate to churn.

3.13 Relationship Between Account Length and Churn

I analyzed the relationship between account length and churn, finding that shorter account durations are associated with higher churn rates.

4: DATA TRANSFORMATION

To prepare the data for modeling, categorical variables such as "state," "international plan," and "voicemail plan" were converted into dummy variables using one-hot encoding.

5: DATA MODELLING

I separated the target variable 'churn' from the features and used the remaining columns as features for model training.

I applied the MinMaxScaler to scale the features to a range between 0 and 1, ensuring that all features contribute equally to the model.

The dataset was split into training and testing sets, with 20% used for testing. I created and trained a Logistic Regression model, then fitted it on the training data.

6: MODEL EVALUATION METRICS

Accuracy: Our classifier shows that our model is 85% accurate.

Precision: Out of all the instances predicted as positive, approximately 52.94% were actually positive.

Recall: The model identified 17.82% of actual positive instances correctly.

F1 Score: The F1-score is 26.67%, balancing precision and recall.

AUC: The AUC is approximately 0.575, suggesting that the model's ability to distinguish between classes is similar to random guessing.

7: CONCLUSION

The churn prediction analysis successfully aimed to build a reliable classifier for SyriaTel. Random Forest was the most effective model for churn prediction, outperforming Logistic Regression and Decision Trees.

The key features impacting churn were total day minutes, customer service calls, and international plan subscription. These insights can help SyriaTel design proactive retention strategies targeted at high-risk customers.

8: RECOMMENDATIONS

I recommend the following actions:

Improve Call Quality: Focus on upgrading infrastructure to enhance call quality.

Enhance Customer Service: Minimize response times and streamline issue resolution.

Offer Tailored Plans for International Subscribers: Develop compelling offers to improve satisfaction.

Implement Proactive Retention Strategies: Use targeted promotions and loyalty programs.

Conduct Regular Analysis: Continuously monitor customer behavior and churn patterns.

9: Next steps

- **Improve Model Accuracy**: Test different models like Gradient Boosting or K-Nearest Neighbors to see if they perform better.
- **Handle Imbalanced Data:** Use oversampling techniques like SMOTE or undersampling to balance the dataset.
- **Add New Features:** Create simple features like the total charge for all calls to see if it improves the model.
- **Monitor Model Performance:** Check the model's accuracy regularly with new data to ensure it still performs well.

THANK YOU