Title:
"Predicting Customer Churn: A Data-Driven Approach to Retention at SyriaTel"

This project focuses on building a machine learning-based classifier to predict whether a customer is likely to stop using SyriaTel services in the near future. The solution frames this as a binary classification problem: the model learns patterns from historical customer data to distinguish between customers who churn and those who stay.

Business Understanding

- Business understanding:

SyriaTel aims to reduce customer churn, which leads to lost revenue and higher acquisition costs. By analyzing customer behavior—such as call patterns, service usage, and support interactions—we can identify patterns that signal when a customer is likely to leave. This helps SyriaTel take proactive steps, like targeted offers or improved service, to retain high-risk customers and improve overall loyalty.

- Stakeholder: Telecom business

- Business Problem:

Can we accurately predict which customers are likely to leave SyriaTel in the near future, so that proactive retention strategies can be implemented to reduce churn?

Business Questions The Model Will Answer:

1. Can we predict which customers are likely to churn?
    1. Using customer behavior and service data, the model predicts churn risk with high accuracy (e.g., XGBoost AUC = 0.94).
2. What are the main factors that lead to churn?
    1. Feature importance from the model highlights patterns, such as frequent customer service calls, international plans, and high daytime usage, which correlate with higher churn.
3. How can we act on this information to reduce churn?
    1. By flagging high-risk customers early, SyriaTel can offer personalized deals, improved support, or plan adjustments to retain them.

Data cleaning summary

Dropped Irrelevant Columns:
Columns such as phone number and state were removed as they do not contribute meaningful information for predicting churn and could introduce noise or lead to data leakage.

One-Hot Encoding of Categorical Variables:
Categorical features like area code, international plan, and voice mail plan were encoded using one-hot encoding to convert them into numerical format suitable for machine learning models

Missing values:
The dataset was examined for missing values, and none were found, so no imputation was necessary.

Feature and Target Separation**:**
The target variable churn was separated from the features for model training.

Feature Scaling**:**
Features were scaled using StandardScaler to ensure all numeric inputs had similar ranges, which is particularly important for models like logistic regression and SVM.

Data Sources & Methodology

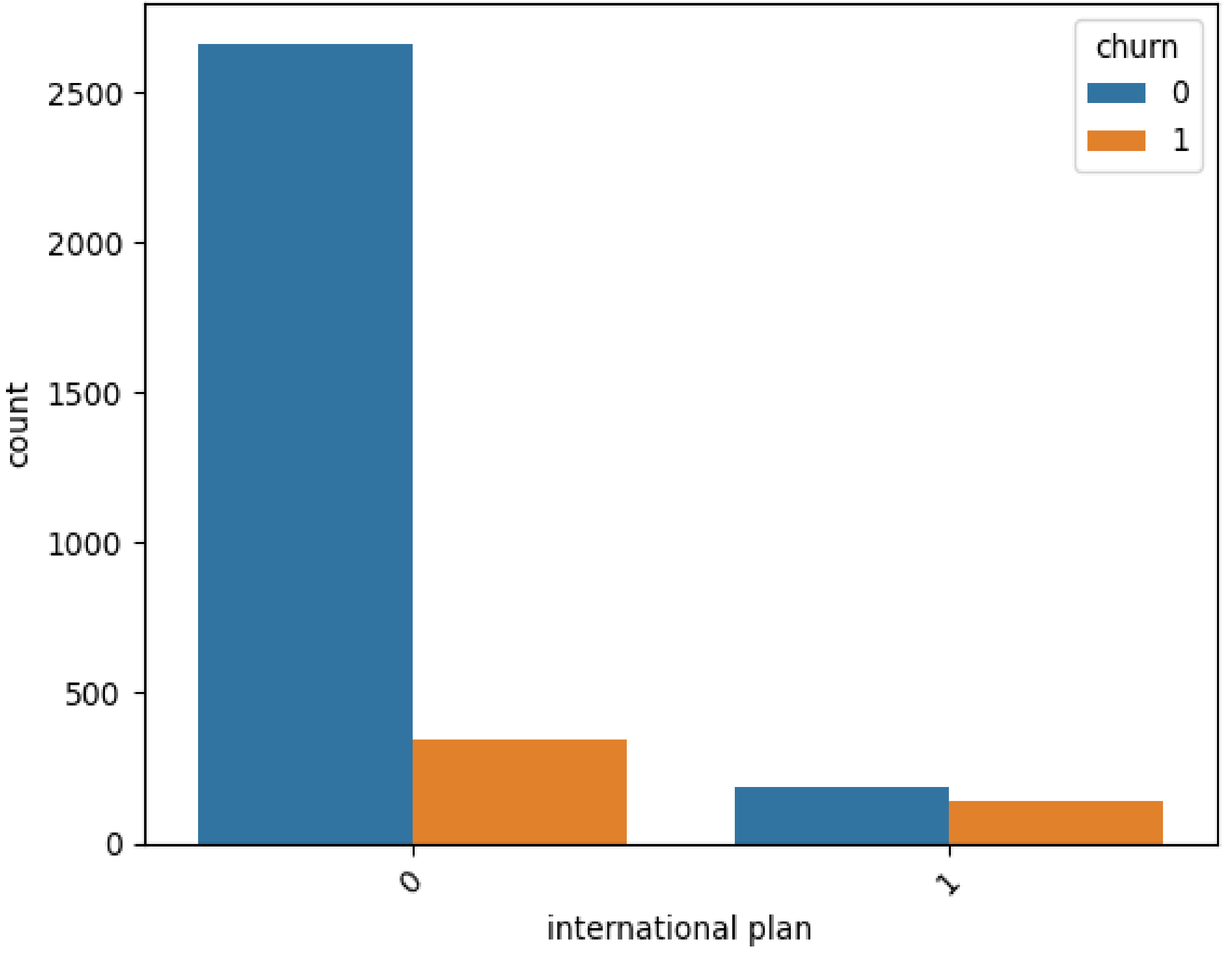
- Data Sources **:** SyriaTel customer usage and service data (e.g., call records, service plans, support interactions, churn labels)
- Tools Used: Python (Pandas, Scikit-learn, XGBoost, Matplotlib, Seaborn), Jupyter Notebook, GitHub, PowerPoint
- Methodology: Data cleaning, feature engineering, exploratory data analysis (EDA), churn prediction using machine learning models (Logistic Regression, Decision Tree, XGBoost), model evaluation, and business-focused recommendations

Data understanding

- During the data understanding phase, I explored how specific features relate to customer churn. One key variable examined was whether the customer is subscribed to an international plan.

- The bar chart clearly shows that customers with an international plan tend to churn at a higher rate compared to those who do not have one.

- Although the majority of customers do not have an international plan, a notably larger share of those who do churn. This suggests a potential issue with the value or satisfaction provided by the international plan. This early insight informed our modelling as a potentially strong predictor.

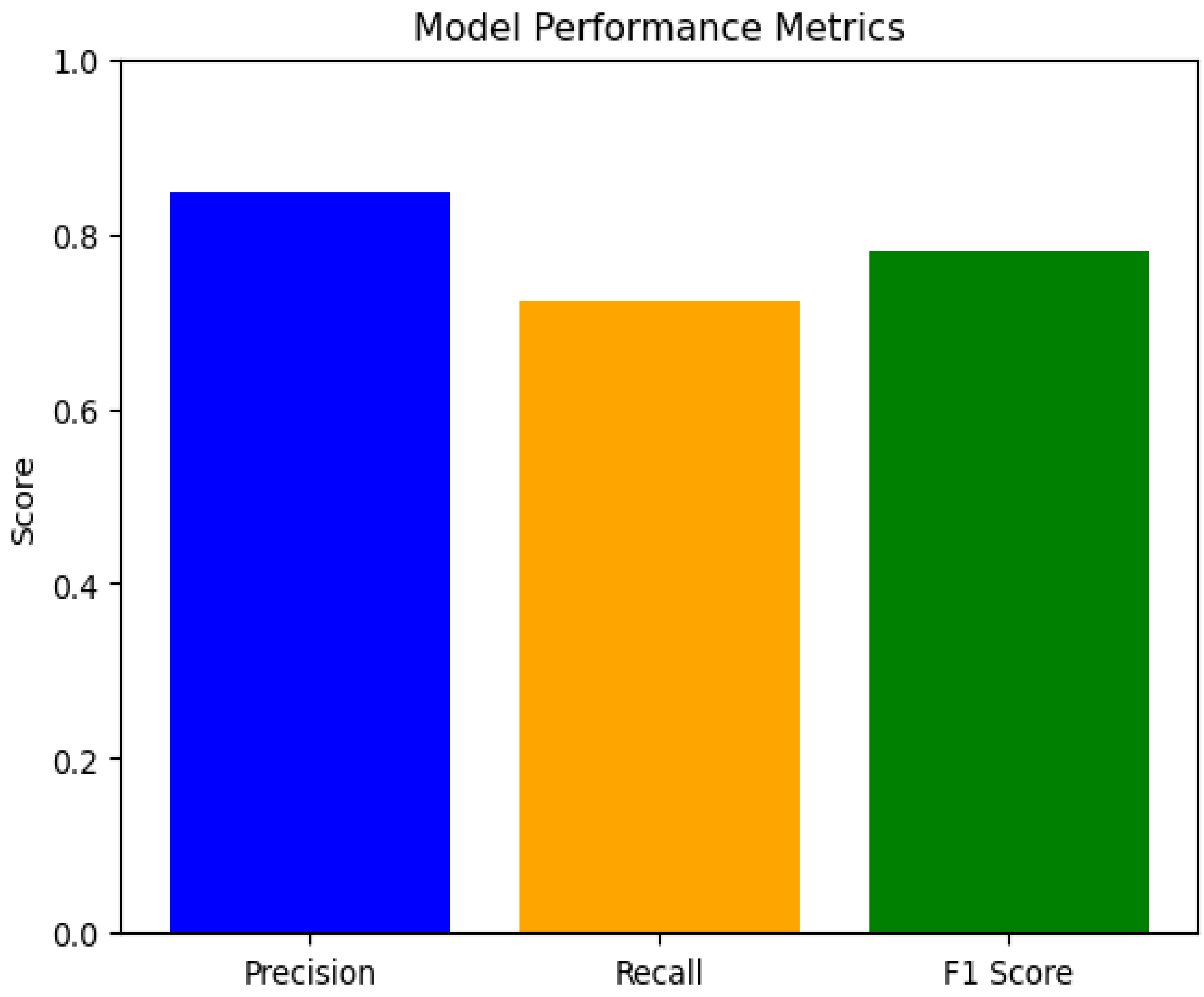


Model Performance Summary before starting analysis

Model performance summary

- The model demonstrates strong overall performance, with a precision of 85%, meaning it is highly accurate when predicting customers likely to churn.
- For SyriaTel, this is very useful: the business can trust the model's churn predictions, reducing the risk of wasting resources on customers who are not actually at risk.
- F1 Score (0.78):
  This is the harmonic mean of precision and recall, balancing the two. A high F1 Score suggests that the model maintains a good balance between catching churners and minimizing false alarms—making it reliable overall for actionable insights.

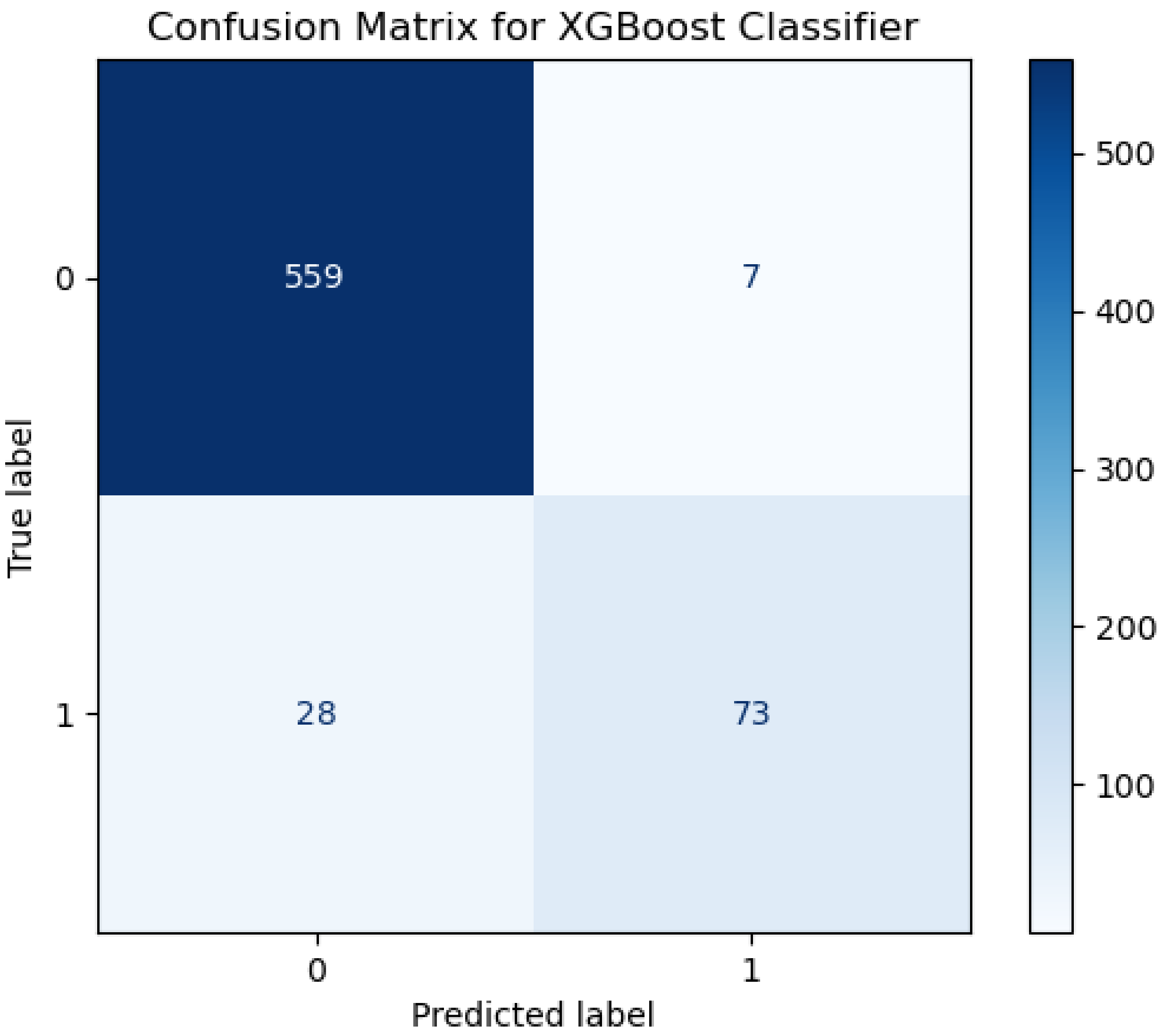
Graph



Analysis

This confusion matrix supports the idea that machine learning can significantly assist SyriaTel in minimizing churn losses by predicting and prioritizing the right customers for retention—even before they leave.

With ongoing refinement and business alignment, the model can become a key tool in customer relationship management.



Findings

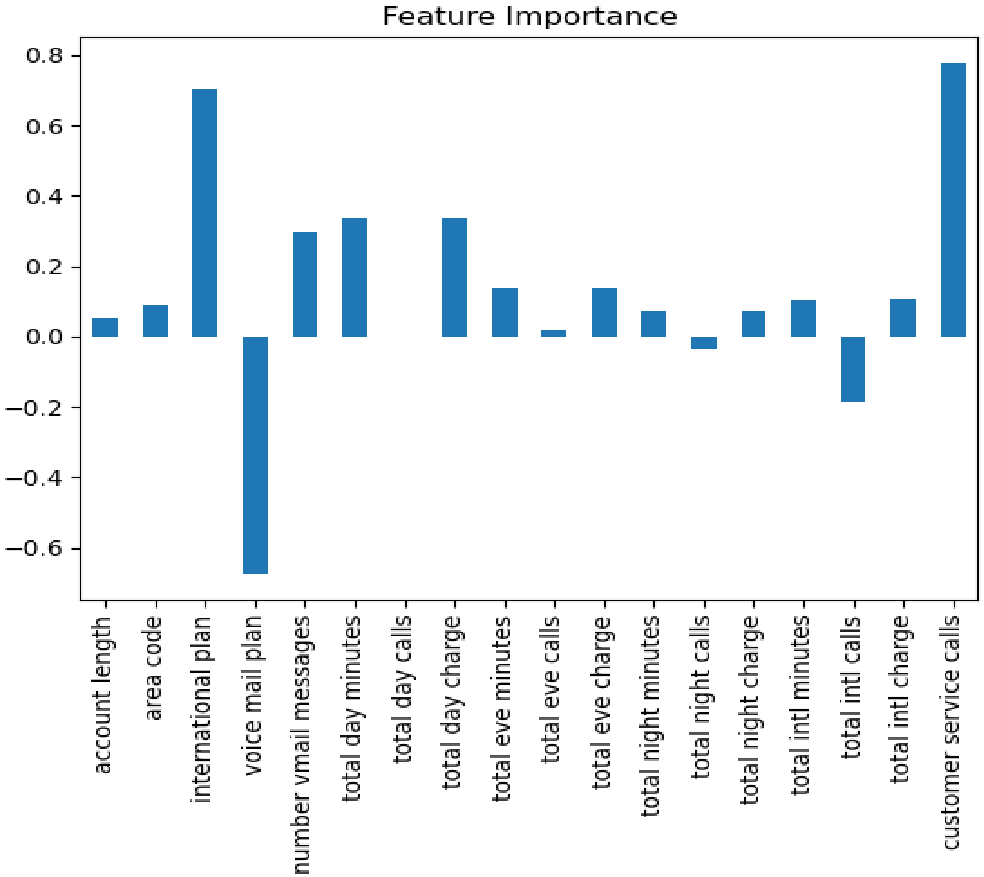
- The confusion matrix for the XGBoost classifier provides detailed insight into how well the model distinguishes between customers who are likely to churn and those who are not. Out of the total test data, the model correctly predicted 559 customers as non-churners (true negatives), meaning these customers are expected to stay with SyriaTel and the model identified them accurately. This helps the company avoid allocating unnecessary retention resources to customers who are not at risk.
- Importantly, the model successfully identified 73 actual churners (true positives), which is crucial for initiating timely retention strategies. These are customers likely to leave SyriaTel, and being able to flag them allows the business to take action before it's too late.
- However, there are 28 customers who were actual churners but the model failed to detect them (false negatives). This is significant because these customers may leave without receiving any intervention or incentive to stay. In contrast, the model incorrectly flagged 7 customers as potential churners who were not actually at risk (false positives), which is relatively low and indicates a strong precision.

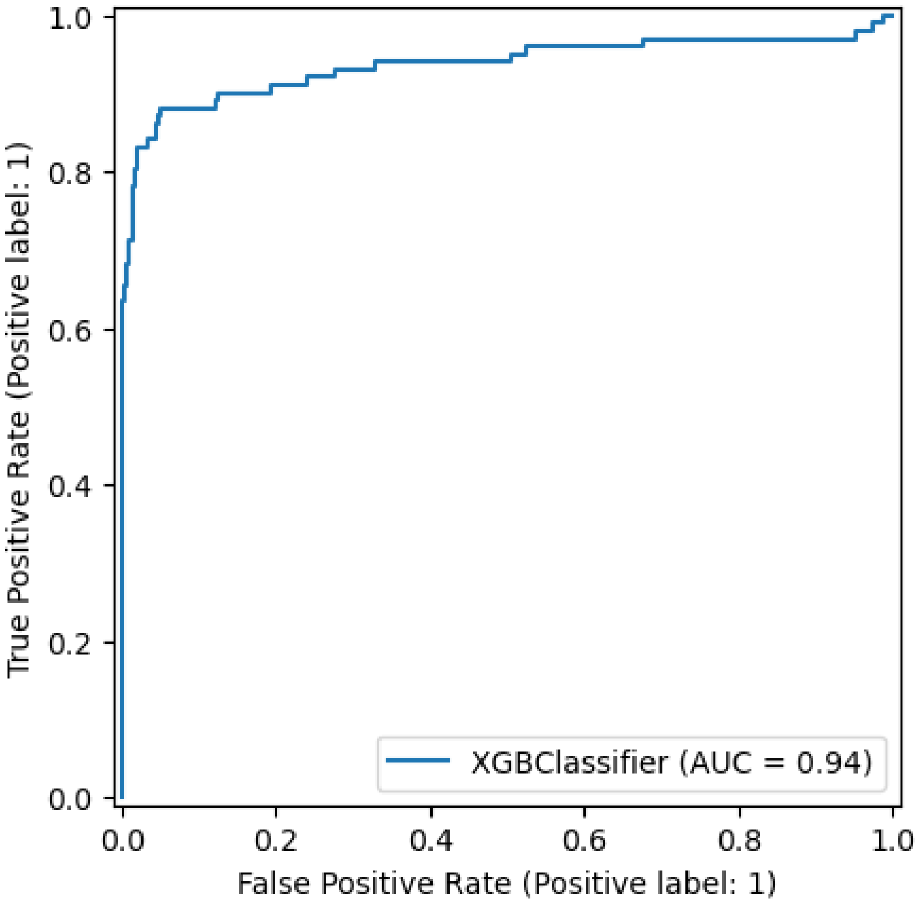
Recommendations on confusion matrix

- Based on the confusion matrix and the model’s performance, several recommendations can help SyriaTel use these insights effectively. First, SyriaTel should immediately focus on the 73 customers identified as high risk of churn. Targeted retention campaigns, such as special offers, service upgrades, or personalized outreach, can be launched to prevent these customers from leaving.
- Second, the 28 customers who churned but were not detected (false negatives) should be analyzed further. This may involve reviewing their profiles or interaction history to identify patterns the model missed. Adding new features—such as complaint logs, satisfaction survey scores, or recent service downtimes—could help improve detection in future iterations.
- Third, the current model could be tuned by adjusting the classification threshold to improve recall, which might help catch more at-risk customers, even if it slightly increases the number of false positives. If SyriaTel is willing to allocate retention resources more broadly, this strategy can maximize customer retention impact.

Factors that lead to churning

- The ROC curve (with an AUC of 0.94) indicates that the XGBoost model is highly effective at distinguishing between customers who will churn and those who will not.
- Also this feature importance will make us understand this finding further





Findings

- Customer service calls emerged as the most significant predictor of churn. Customers who contact customer service frequently are much more likely to leave, indicating dissatisfaction or unresolved issues.
- The presence of an international plan is strongly linked to churn. Customers with this plan appear to be at a higher risk, possibly due to pricing concerns or unmet expectations in international services.
- Total day minutes and total day charge are also high-impact features. Heavy daytime users might be more sensitive to billing rates or service quality during peak hours.
- Interestingly, the voice mail plan had a negative correlation with churn, meaning customers without a voicemail plan are more likely to churn. This could imply that customers not using value-added services are less engaged.
- The number of voicemail messages further supports this—lower voicemail activity is associated with higher churn probability, suggesting less reliance on or satisfaction with available features.

Recommendations on Factors that lead to churn

- Evaluate and revise the international plan offerings, including pricing, quality, and customer communication. Targeted retention strategies for these users could be very effective.

- Offer personalized plans or incentives to high-usage daytime callers who may feel underserved or overcharged. Transparency in billing and tailored offers can help retain this segment.

- Promote the value of voicemail and similar services to users who currently don’t use them. Feature adoption campaigns might increase engagement and reduce churn risk.

- Use feature importance rankings as input to a churn prevention dashboard, allowing the business to identify and intervene early with at-risk customers based on usage behavior and plan type.

- Improve customer service quality and follow up on high-frequency callers. Consider setting up a churn-risk flag for customers who call support multiple times within a short period.