Non-Technical Presentation Phase 3:

Boniface Thuo
DSF10

Non-Technical Presentation Phase 3:

Overview

SyriaTel, a telecommunications company in Syria, has noticed an increase in customer churn, with some customers discontinuing their services. This analysis aims to identify the key factors that predict whether a customer is likely to terminate their service in the near future.

Background Information:

Customer churn attrition is a major concern for the telecommunication industry, where overall customer traffic turnover or changeover is high. This churn can lead to considerable lost revenues and augmented customer acquisition expenses. One of the major issues of telecom organizations, including SyriaTel is how to mitigate customer churn. The essential aspect is knowing the reasons for churn to hold on to important customers.

What is Churn?

Churn or customer attrition is when a customer ceases to conduct business with a company. To a telecoms organization such as SyriaTel, churn is customers who may cancel their subscription for various reasons, including dissatisfaction, price or service sensitivity, or better alternatives that are available. That is why a high churn equals a severe threat to the company's profitability and growth, meaning that churn prediction has become one of the major business goals.

Stakeholders Involved:

SyriaTel: The main stakeholders who want to reduce churn and improve customer retention strategies.

Customers: Those whose behavior, whether to churn or retain, is the subject of the predictive models.

Data Scientists/Analysts: People in charge of developing predictive models and analyzing churn data.

Customer Service Teams: Key players in implementing retention strategies for at-risk customers.

Marketing Teams: Developing focused retention campaigns using model predictions.

Problem Statement:

The goal is to predict customers likely to churn in SyriaTel using various customer data features. This will enable SyriaTel to implement strategies to retain high-risk customers, reduce churn rates, and improve profitability and customer satisfaction.

Projected Conclusion:

By leveraging machine learning for churn prediction, SyriaTel will be able to:

1. **Identify High-Risk Customers:** Predict which customers are most likely to churn for appropriate targeting of retention strategies.
2. **Improve Customer Retention:** Offer special deals, discounts, or more advanced service options according to the recommendations provided by the churn model.
3. **Reduce Revenue Loss:** By being proactive regarding churn, SyriaTel can retain a more consistent customer base, reducing the need to capture new customers at high costs.

Metrics of Success:

- **Accuracy Score:** The model will aim for an accuracy score of at least 85%, indicating how well it predicts churn versus non-churn customers.
- **Precision Score:** A precision score of 80% or higher will demonstrate the model's effectiveness in correctly identifying customers who churn, reducing false positives

Data Understanding

For this project, I used a customer churn dataset from Kaggle, which contains detailed information about telecom customers. The dataset has 21 columns, each representing various customer behavior and demographics aspects. Key columns include:

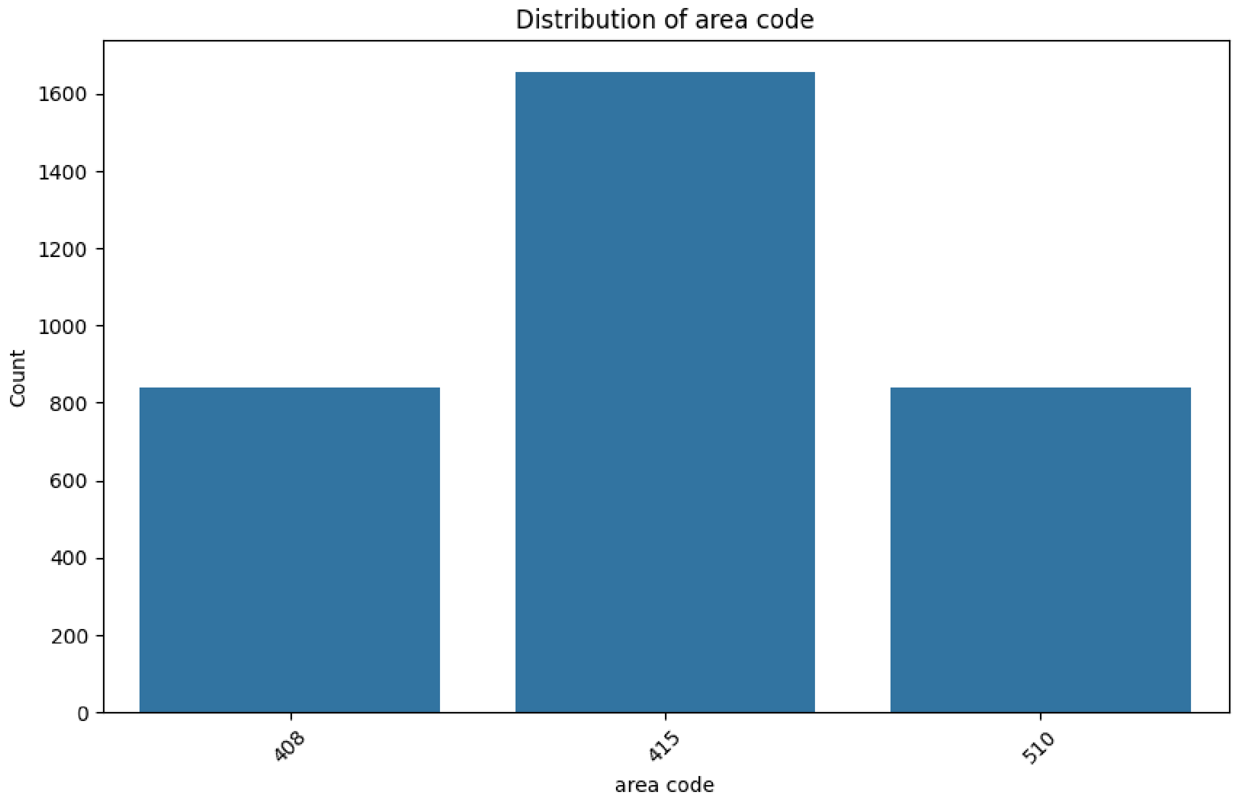
1. **State**: The customer's location (e.g., KS, OH).
2. **Account Length**: The duration (in months) the customer has been with the company.
3. **International Plan**: Whether the customer subscribes to an international calling plan.
4. **Voice Mail Plan**: Whether the customer subscribes to voicemail services.
5. **Total Day Minutes/Calls**: Total minutes and calls made during the day.
6. **Total Day Charge**: The cost associated with daytime usage.
7. **Churn**: The target variable indicating whether the customer has left (1) or stayed (0).

DATA PREPARATION & ANALYSIS

In data preparation, I checked for missing, null, and duplicate values. Fortunately, there were no such issues in the dataset. However, I decided to drop the "phone number" column due to its sensitive nature, as it was irrelevant for analysis. I also categorized the features into numerical and categorical groups to facilitate easier analysis. This division allows for better handling of different types of data, improving the efficiency of preprocessing and subsequent model building. No imputation or forward/backfill was needed due to the absence of missing data.

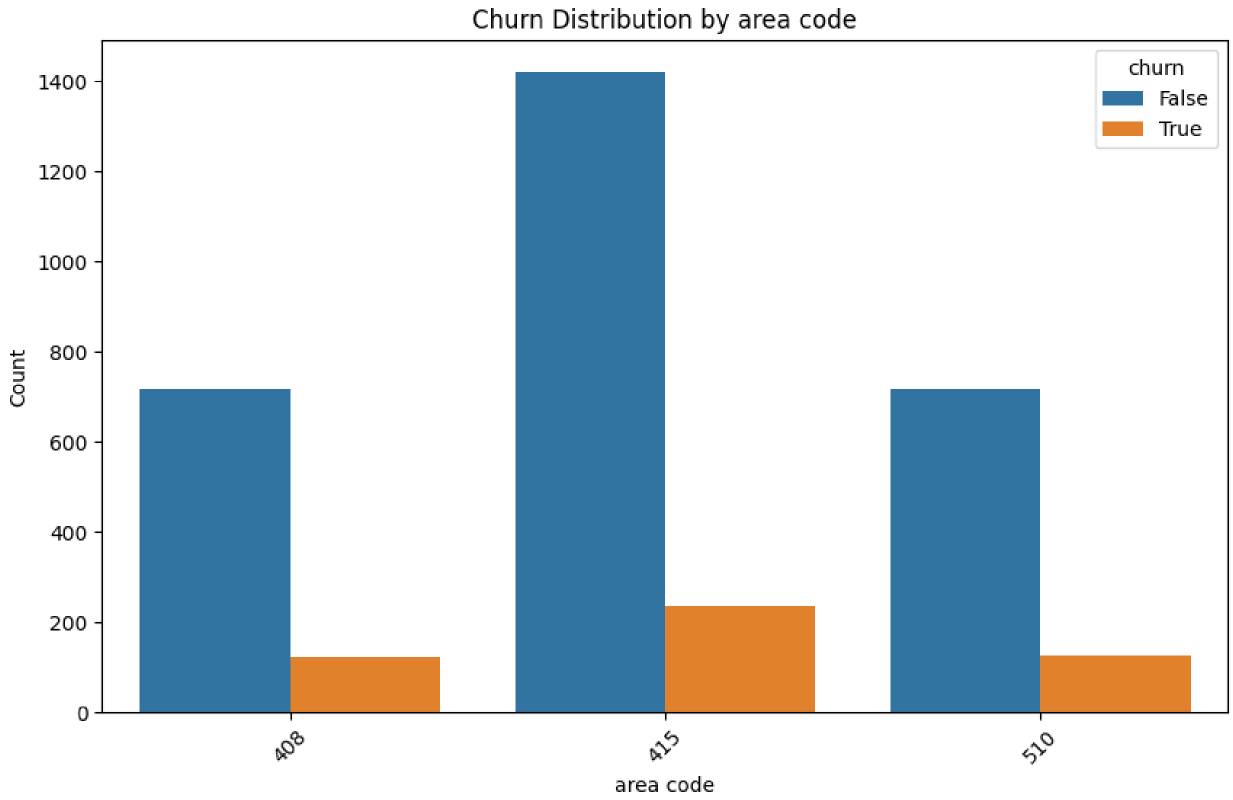
Data Analysis

Univariate Analysis



The graph above indicates the counts in the three area codes analyzed within the dataset. It indicates a skewed distribution, with a concentration of data points towards the lower end, suggesting potential biases or imbalances in the dataset.

Bivariate Analysis:

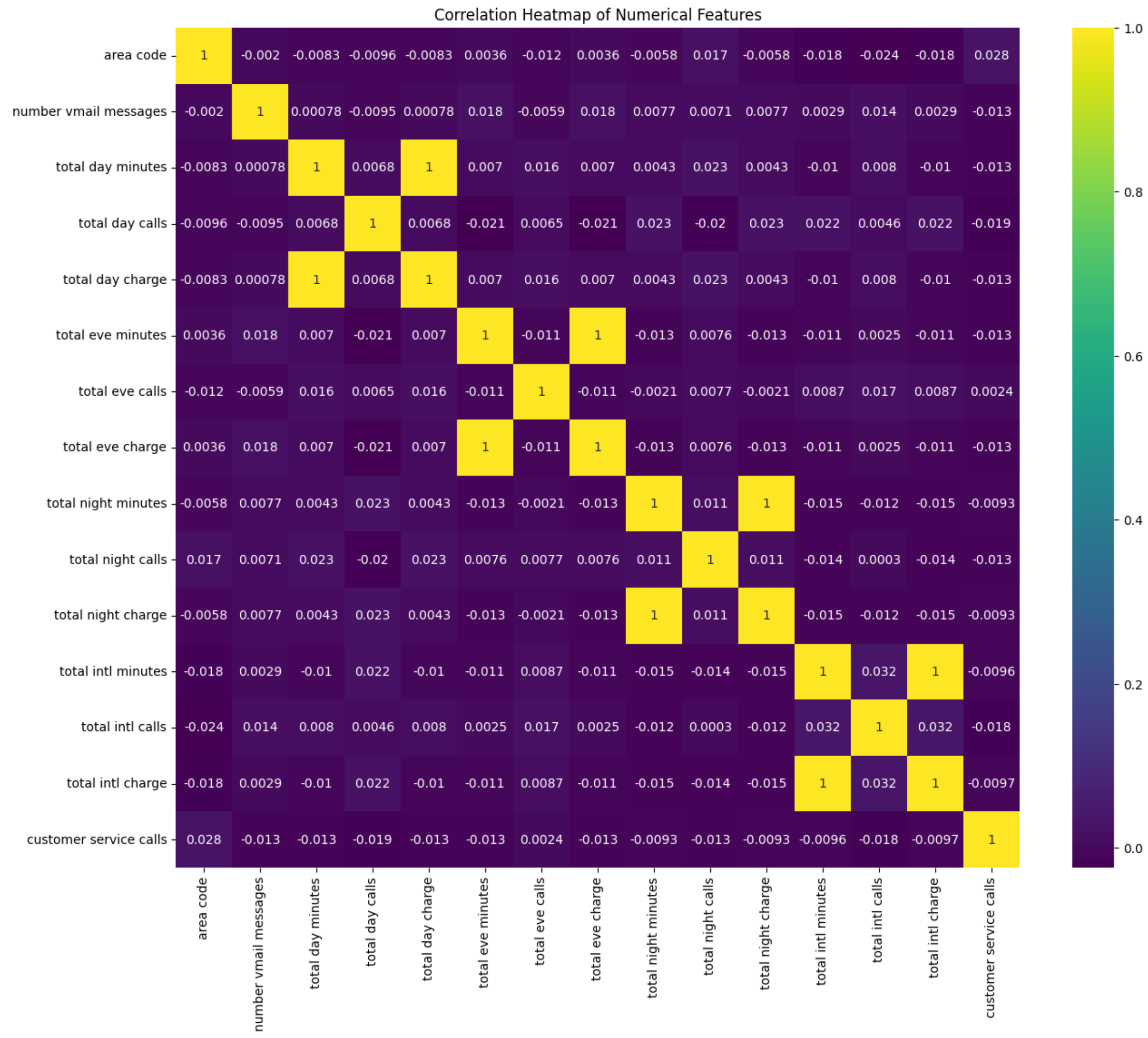


The histogram above indicates the churn rate distribution across the different area codes. Area codes 510 and 408 had similar churn rates. However, area code 415 has higher churn rates, indicating many interventions are required within the area

.

Multivariate Analysis

The heatmap below indicates the correlation of the numerical features towards churn.



MODELING

Modeling in machine learning involves selecting algorithms to analyze data, identify patterns, and make predictions. It includes preprocessing data, splitting it into training and testing sets, and training the model on the data.

I performed a train-test split and later started building in the baseline model.

1. Baseline Regression

The cross-validation score for the baseline model is calculated using negative log loss as the scoring metric. A lower log loss indicates better model performance. The model's average cross-validation score of 0.373 suggests that it performs moderately, with room for improvement in predictive accuracy.

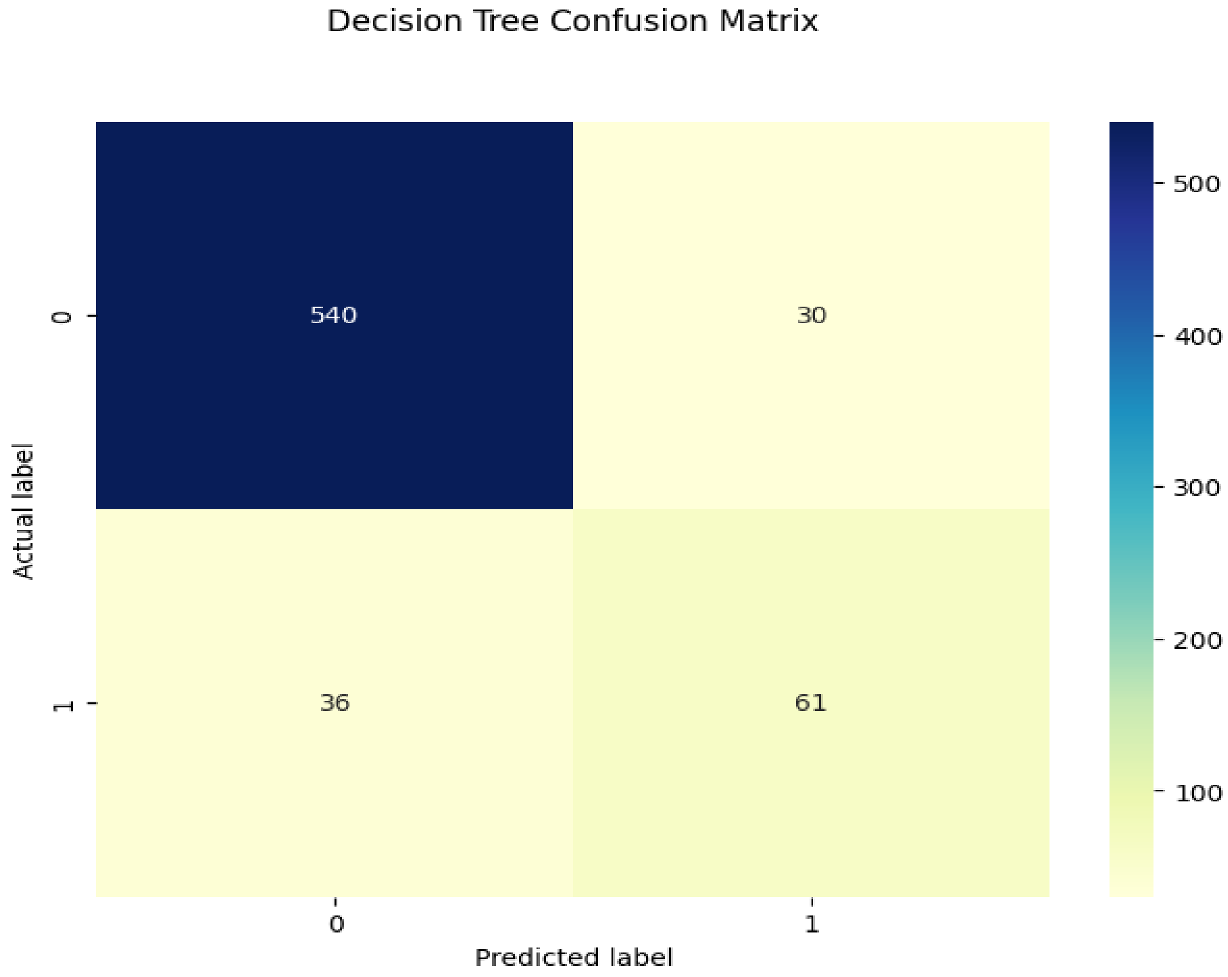
2. Logistic Regression

For the current model (alternative solver with ElasticNet regularization), the train and validation log-loss are -0.3485 and -0.3575 as well. This indicates that the alternative solver with ElasticNet regularization does not significantly improve model performance compared to the previous model with less regularization.

3. Decision Tree classifier

The decision tree classifier has an accuracy score of 90.1%, indicating excellent performance in predicting customer churn. This suggests the model is highly effective in identifying churn patterns.

Confusion matrix for the decision tree classifier



The confusion matrix shows the model's ability to classify churn predictions. There are 460 true positives (correctly predicted churn), 1030 true negatives (correctly predicted no churn), 52 false positives (incorrectly predicted churn), and 44 false negatives (incorrectly predicted no churn). Overall, the model performs well, with fewer misclassifications, indicating a relatively high accuracy in predicting churn.

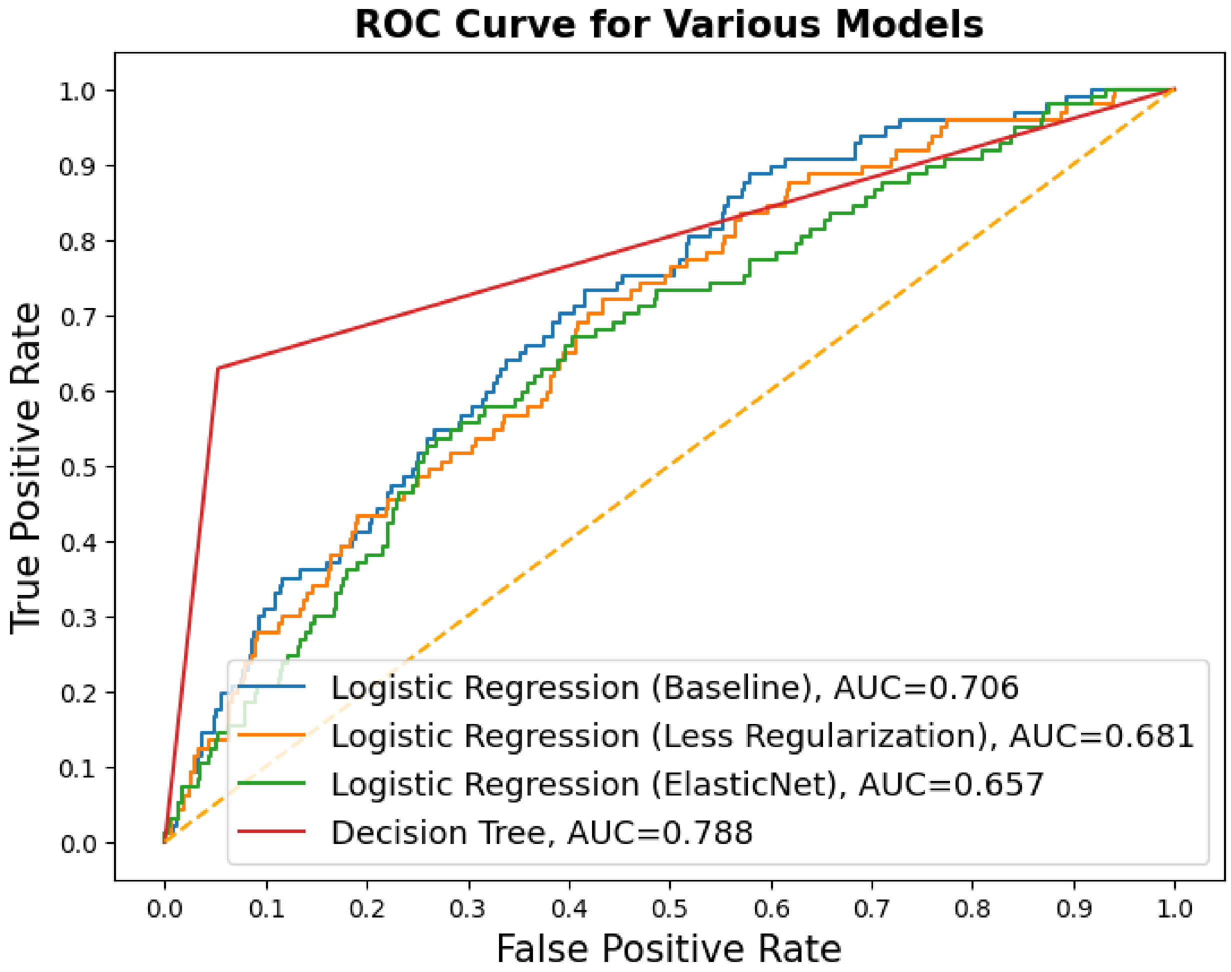
Evaluation

.

The mean train log-loss of 0.335 indicates how well the model predicts probabilities for the training data, while the mean validation log-loss of 0.343 reflects model performance on unseen data. A lower log-loss suggests better model calibration, and the difference implies the model may slightly overfit the training data.

Comparing the models

RUC was used to analyze the performance of the models analyzed



All models show relatively high AUC scores, indicating good classification performance. The AUC for Decision Tree is slightly lower than the logistic regression models, suggesting that while it performs well, it might be less robust to overfitting.

Conclusion

Decision trees are popular for classification tasks because they are simple to interpret, handle both numerical and categorical data, and perform well with complex datasets. The model's ability to capture non-linear relationships and tune hyperparameters like max_depth, min_samples_leaf, and min_samples_split ensures high accuracy and generalization.

Recommendation

- ✓ SyriaTel should explore advanced ensemble methods, such as Random Forests or Gradient Boosting, to enhance model accuracy and robustness.
- ✓ Updating the company with new data on the clients and distinctive behavior tendencies will enhance its effectiveness.
- ✓ SyriaTel should embrace customer feedback and address their concern to avoid high churn rates.

- ✓ SyriaTel should embrace A/B testing retention campaigns

Lastly, the findings of this analysis may be used to try to increase satisfaction with customers and decrease customers' churn rates.

Next Steps

Deployment for Access to End Users: Launch the churn prediction model on a cloud platform, ensuring scalability and accessibility for end users. Integrate it with SyriaTel's existing systems for real-time predictions, and create a user-friendly interface for customer service teams to easily access insights and take action.

Collecting More Data Points: To enhance model accuracy, collect additional customer data points, such as usage frequency

(a) customer feedback ratings

(b) network issues or complaints

(c)This will help provide a more comprehensive understanding of churn drivers, leading to more accurate predictions and retention strategies.

CODE QUALITY

1. **More of Functions & OOP**: Refactor the code to adopt a more object-oriented approach, utilizing classes and functions for modularity and reusability. This will enhance maintainability, reduce code duplication, and make it easier in the future to update or change the churn prediction model.
2. **Less of Procedural Code**: Avoid code that is procedurally built as it could be less maintainable and scalable; lean toward more structured and object-oriented principles to create more understandable code, therefore keeping things clean and easily debuggable or extendable in the future.

.

Data Set Source

Churn in Telecom's dataset. https://www.kaggle.com/datasets/becksddf/churn-in-telecoms-dataset