SYRIATEL CUSTOMER CHURN CLASSIFICATION ANALYSIS

By Winnie Awino Odoyo

CONTENTS

Problem Statement

SyriaTel, a telecommunications company, is facing a high customer churn rate, negatively impacting its revenue growth and profitability. Customer acquisition is up to five times more expensive than retaining existing customers, making this a significant financial challenge. In an increasingly competitive market with numerous alternatives for customers, high churn rates make it difficult for SyriaTel to maintain its market share, reinvest in innovation, and achieve long-term growth.

Business Understanding

Churn, also known as customer attrition, refers to the percentage of customers who discontinue or stop using a company's services during a given period. For telecommunications companies like SyriaTel, churn is a critical metric because it directly impacts revenue growth and profitability. When a customer leaves, the company not only loses the revenue from that customer but also incurs higher costs in trying to acquire new customers to replace them.

Acquiring new customers is significantly more expensive than retaining existing ones, with estimates suggesting that it's up to 5 times more costly. In a highly competitive global market, telecom companies face the constant challenge of retaining customers who have various options and can easily switch providers. High churn rates increase the costs of acquiring new customers and maintaining market share, affecting a company's ability to grow and invest in new technologies or services. Predicting and reducing churn is not only essential for sustaining SyriaTel's market position but also for ensuring its capacity to grow, innovate, and provide competitive services in a rapidly evolving industry.

Data Understanding

This dataset was obtained from kaggle and contains customers' data from SyriaTel, a telecommunications company. It includes various features related to customer behavior, usage 3 patterns, and service details, which are crucial for analyzing customer churn and understanding the factors that influence retention.

The dataset entails 21 columns and 3333 rows. The cleaning of this dataset entailed dealing with outliers, dealing with duplicates, checking for missing values which the dataset didn't have, harmonizing the various data types and also operating both dummy encoding and label encoding in preparation of the data for modeling.

Objectives

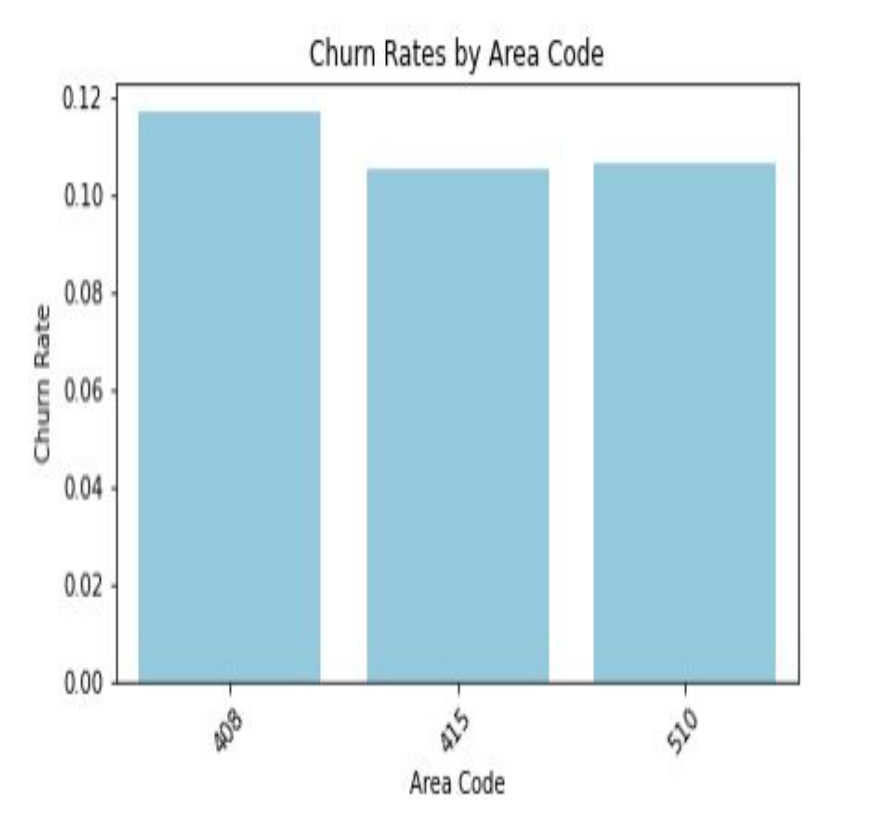
To Identify Churn Rates by Region and which region is highly affected.

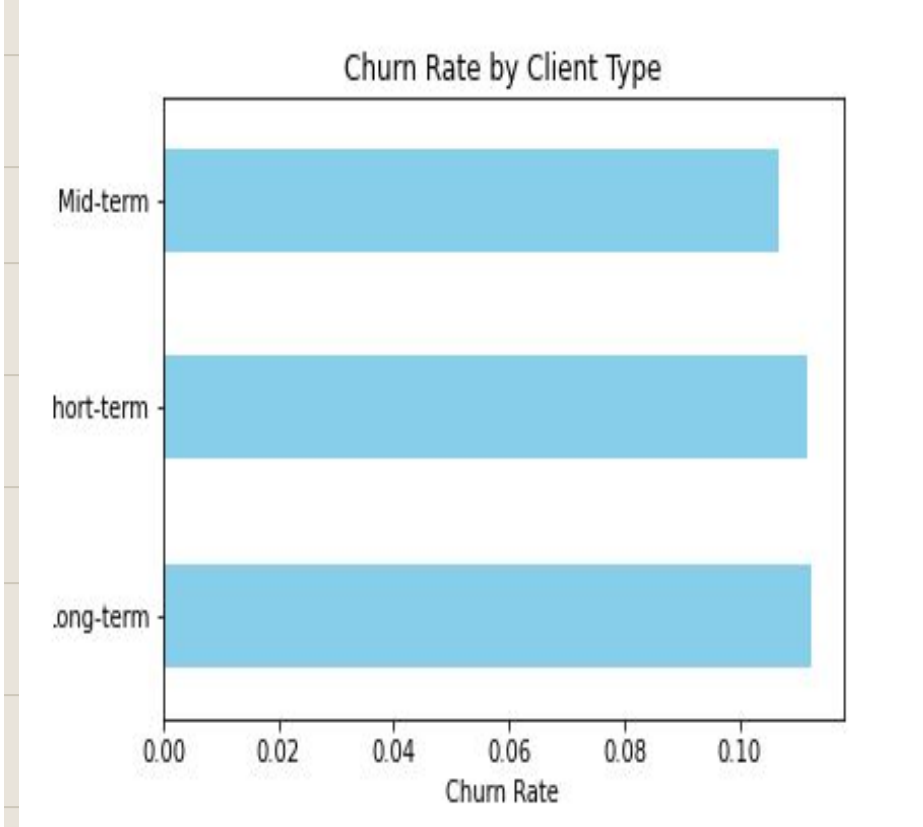
To examine the impact of Account Length on Churn. Which customers are highly prone to attrition.

To assess the influence of subscription plans (voicemail and international plans) on churn.

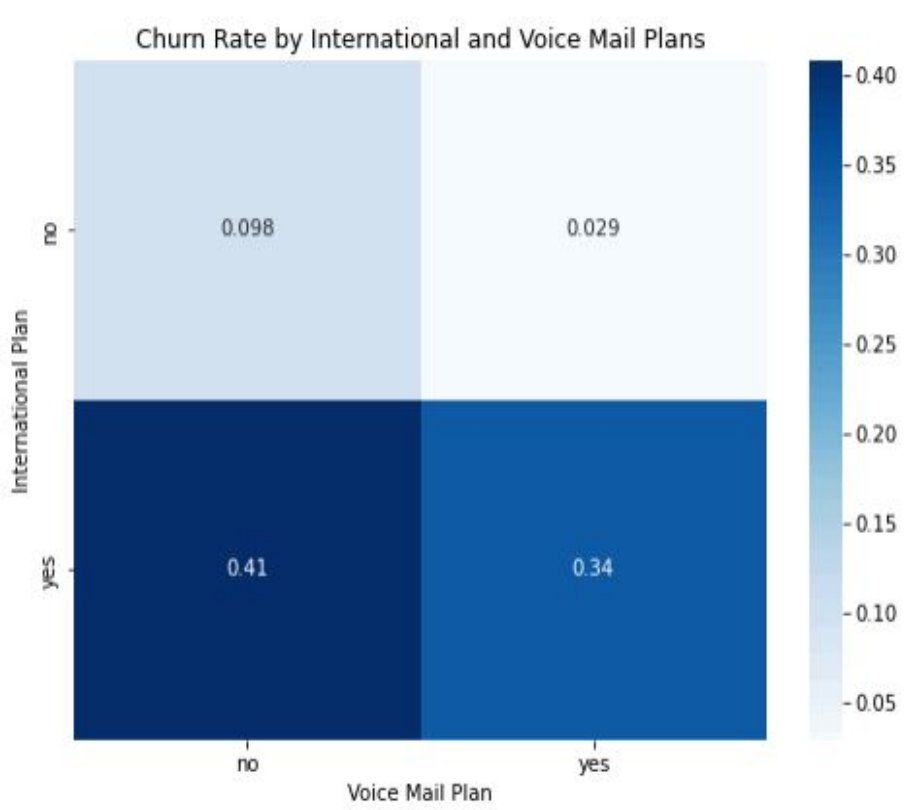
To evaluate the impact of customer service calls on churn and determine if there is a statistically significant relationship.

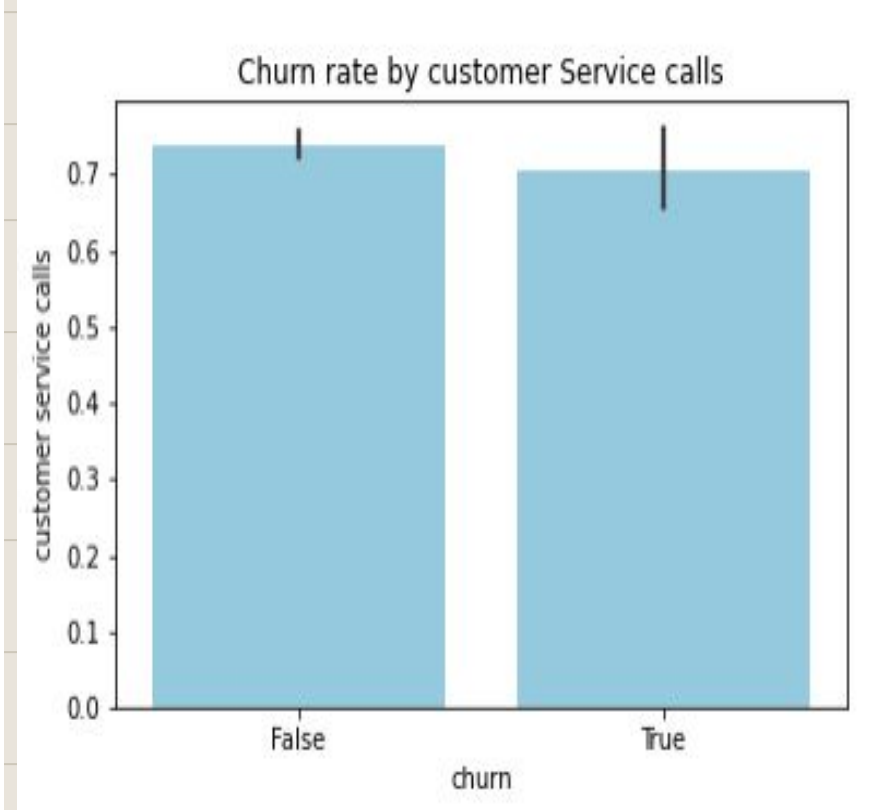
Analysis


Churn Rates by Area Code
Churn Rate
0.12
0.10
0.08
0.06
0.04
0.02
0.00
408
415
510
Area Code


Churn Rate by Client Type
Mid-term
hort-term
.ong-term
0.00
0.02
0.04
0.06
0.08
0.10
Churn Rate

Analysis Cont.





Modeling

This section involved training four distinct models: basic logistic regression and decision tree models, along with enhanced versions of both.

The second models incorporated feature selection, addressed class imbalance using class weights, applied regularization (for logistic regression), and employed hyperparameter tuning (adjusting max depth, min samples and entropy) for the decision tree models. Logistic regression and decision tree models were chosen as they are well-suited for analyzing the categorical nature of the target variable (churn).

Model Evaluation

These models were analysed further using four different metrics entailing, cross validation score, confusion matrix, ROC Curve and classification Report as detailed below.

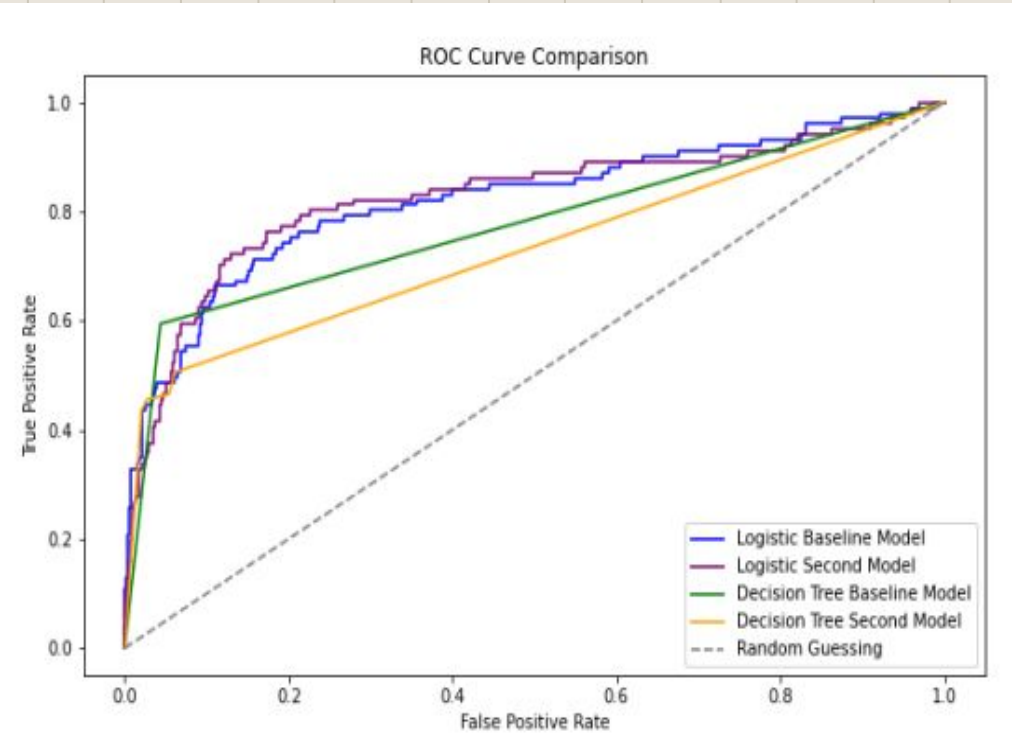
a). Cross-Validation Score

```
Logistic Baseline Model CV Scores:  0.34987804878048784
Second Logistic Model CV Scores:  0.8225609756097562
Baseline Decision Tree Model CV Scores:  0.6596341463414633
Second Decision Tree Model CV Scores:  0.5563414634146342
```

According to this metric, the Second Logistic Model has the highest validation score based on recall, as the primary focus is to identify as many churners as possible. This model outperforms the others by identifying a significantly higher proportion of churners.

Model Evaluation Cont.

b). ROC-Curve



Interpretation

- **Logistic Baseline Model:** Performs well but is outperformed by the logistic second model, showing room for improvement in feature selection or tuning.
- **Logistic Second Model:** The best-performing model, achieving the highest TPR and lowest FPR, indicating effective optimization for churn prediction.
- **Decision Tree Baseline Model:** Provides moderate performance, but its ROC curve suggests it struggles with generalization compared to the logistic models.
- **Decision Tree Second Model:** Shows slight improvement over the baseline decision tree, but still lags behind the logistic models in predictive power.

Model Evaluation Cont.

c). Classification Report
- Logistic Regression Models

According to this metric, this is how the models perform,

- **Logistic Baseline Model** performs well for non-churn customers(high recall and precision for class 0)but has difficulty identifying churn.This suggests a high number of false negatives (failing to predict churn).
- **Logistic Second Model** improves churn prediction compared to the baseline model but at the cost of misclassifying more non-churn customers as churn (lower recall for class 0).
- **Decision Tree Models** (both baseline and second) perform well for non-churn customers, but their performance in detecting churn is still not optimal, with relatively low recall for churn.

Classification Report (logistic Baseline Model):

	precision	recall	f1-score	support
0	0.92	0.98	0.95	738
1	0.72	0.39	0.50	101
accuracy			0.91	839
macro avg	0.82	0.68	0.73	839
weighted avg	0.90	0.91	0.90	839

Classification Report (logistic Second Model):

	precision	recall	f1-score	support
0	0.96	0.80	0.87	738
1	0.34	0.77	0.47	101
accuracy			0.79	839
macro avg	0.65	0.78	0.67	839
weighted avg	0.89	0.79	0.82	839

- **Decision Tree Models**

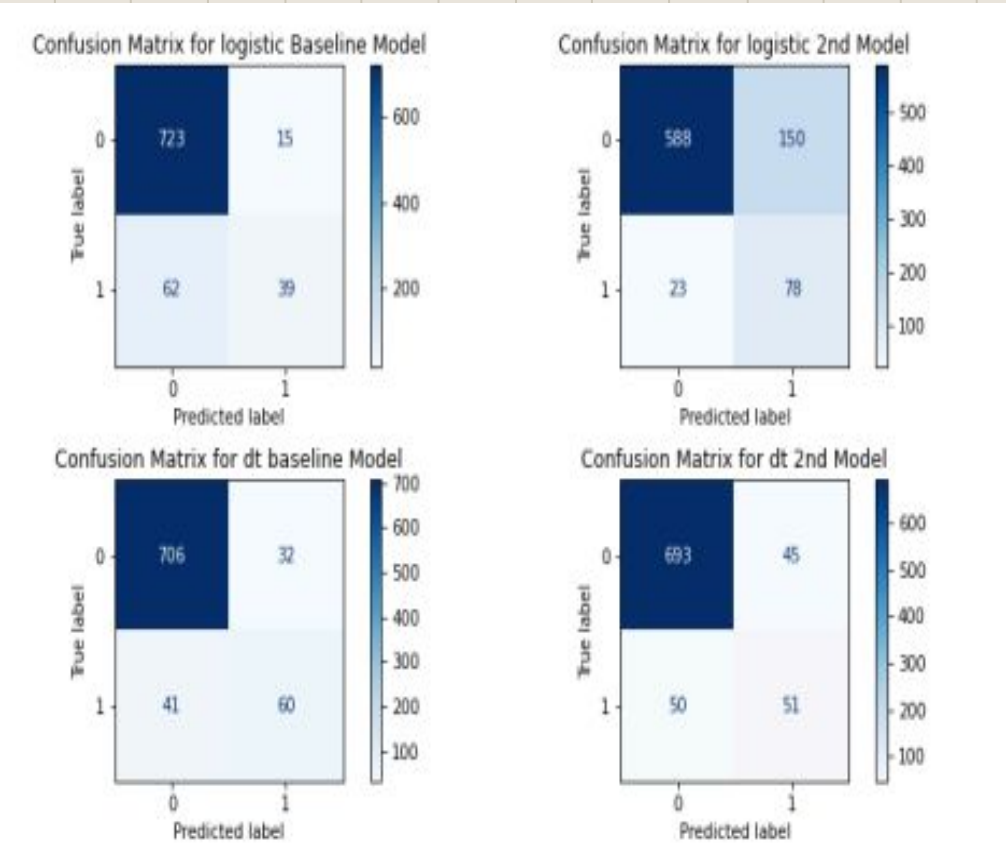
Classification Report (dt Baseline Model):

	precision	recall	f1-score	support
0	0.95	0.96	0.95	738
1	0.65	0.59	0.62	101
accuracy			0.91	839
macro avg	0.80	0.78	0.79	839
weighted avg	0.91	0.91	0.91	839

Classification Report (dt Second Model):

	precision	recall	f1-score	support
0	0.93	0.94	0.94	738
1	0.53	0.50	0.52	101
accuracy			0.89	839
macro avg	0.73	0.72	0.73	839
weighted avg	0.88	0.89	0.89	839

Model Evaluation Cont.



d).Confusion-Matrix Interpretation

Logistic Regression Models:

-The second logistic model has a higher number of True Positives (78) but also a higher False Positive rate (150). This indicates that it may incorrectly classify a higher number of non-churn customers as churn, which could be problematic, depending on the cost of false positives. The first logistic model has fewer false positives and false negatives, but its True Positives (39) are lower, which suggests it may be under-predicting churn.

Decision Tree Models: -Both decision tree models (baseline and second model) perform similarly, with a relatively higher True Positive (60) rate and fewer False Positives compared to logistic regression. The decision tree models seem to strike a better balance between identifying churn customers correctly and avoiding false positives.

Model Evaluation Cont.

Final Model

The second logistic regression model emerged as the best-performing model across multiple evaluation metrics. It achieved the highest validation score, indicating robust overall performance. On the ROC curve, it demonstrated effective optimization for churn prediction, with the highest True Positive Rate (TPR) and lowest False Positive Rate (FPR). The confusion matrix highlighted 10 its strength in correctly predicting churn with the highest number of true positives, although it also showed a tendency to overpredict with increased false positives.

Additionally, the classification report emphasized its superior recall, reflecting the model's strong ability to identify churn cases accurately Based on the analysis of the models, the second logistic regression model is the best choice for predicting customer churn at Syria Tel. While it has a slightly higher False Positive rate, its superior recall score indicates that it is more effective at identifying churners, which is the primary goal for this business problem. By capturing more churners (True Positives), the model offers a better chance to retain high-risk customers, which is more cost-effective than failing to identify churners and losing them to competitors.

Although the higher False Positive rate poses a risk of misclassifying non-churn customers as churners, the financial cost of reaching out to false positives is generally lower than the cost of not addressing actual churners. This trade-off makes this model more suitable for improving customer retention, which is crucial for SyriaTel's profitability and long-term growth. This final model achieves the stated objective by optimizing the most critical features influencing churn prediction,such as account details, usage metrics, customer service interactions, and geographic information. Given its potential, it should be prioritized for further tuning to optimize its accuracy and enhance customer retention strategies.

Conclusion

- The highly affected region for churn is 408, which has the leading churn rate

- The churn rate analysis based on account length shows that long-term customers (over 14 years) have the highest churn rate, followed by short-term customers (7 years or less) with the second-highest churn rate, while mid-term customers (7-14 years) exhibit the lowest churn rate.

- The findings reveal that customers with an International Plan but no Voicemail Plan have the highest churn rate, followed by those with both plans, while customers without either plan have the third-highest churn rate, and those with only a Voicemail Plan have the lowest churn rate.

- The analysis indicates that the average number of customer service calls is relatively high for both churners and non-churners, with no statistically significant relationship between customer service calls and churn rate.

- Based on the analysis, the second logistic regression model stands out as the most effective for predicting customer churn at SyriaTel.It meets the success criteria achieving a 77% recall. While it has a slightly higher False Positive rate, its superior recall score ensures it effectively identifies churners, a critical aspect for enhancing retention strategies.

The model highlights key factors influencing churn, including call charges, account length, geographic region, subscription options, and the number of calls/minutes relative to charges.

Recommendations

- SyriaTel should conduct further research to identify the factors driving the high churn rate in the 408 region, which differs from others, and develop targeted strategies for improving customer retention.

- SyriaTel should focus on retaining long-term customers by addressing potential dissatisfaction and innovating offerings, while improving short-term customer satisfaction through better onboarding and support. Mid-term customers should be engaged with loyalty programs to maintain their retention.

- SyriaTel should focus on retaining International Plan customers, especially those without a Voicemail Plan, by offering cost-effective strategies like enhancing plan benefits, bundling services, and improving customer support.

- Since customer service calls are not significantly related to churn rate, S SyriaTel should shift focus to improving other key aspects of the customer experience, such as product quality, pricing, and service offerings, to enhance customer retention and reduce churn.

- Based on the findings, SyriaTel should prioritize further tuning and optimization of the second logistic regression model, as its superior recall score makes it more effective at identifying churners, thereby enhancing customer retention strategies and reducing churn.

NEXT STEPS

Deploy the logistic regression model into SyriaTel customer retention system. Integrate the model's predictions into real-time operations, such as CRM systems, to trigger automated retention actions like personalized offers or loyalty programs for at-risk customers.The model should also be continuously updated with new customer data to maintain its predictive power and accuracy over time.

There is a need for a deep dive into the 408 region to gather customer feedback or detailed insights into regional challenges. Further qualitative analysis may be needed to understand if factors like service quality, local competition, or pricing are impacting retention.

Implement a feedback loop to gather insights from customers who churn and those who stay, continuously improving the customer experience and identifying pain points that were not previously addressed. Additionally, a structured process is needed for collecting and acting on customer feedback.

THANK YOU!