Objective: The primary objective of this analysis is to predict customer churn—whether a customer will leave or remain with the company. By identifying the patterns that lead to churn, we aim to provide actionable insights that allow the business to take proactive measures to retain valuable customers. Understanding the key factors that influence customer decisions will enable the development of a predictive model, which will contribute to improving customer retention strategies.

Description of the Dataset: The dataset contains customer account details, call behavior metrics (day, evening, night, and international call minutes), and the number of interactions customers have with customer service. The goal of the analysis is to identify which factors contribute most to customer churn, uncover patterns in the data, and use these insights to build an effective predictive model. By understanding these contributing factors, the business can better target its retention efforts.

Exploratory Data Analysis (EDA): During the exploratory analysis, the following key insights were observed:

1. **Relationships between charges and minutes**: There is a perfect correlation between the charge columns (e.g., day charge, evening charge) and the corresponding minutes columns (e.g., day minutes, evening minutes), while there is no significant relationship between these charge and minute variables and other columns in the dataset.
2. **Customer Service Interactions**: Customers who contact customer service more frequently tend to have a higher likelihood of churning. This is likely due to dissatisfaction or unresolved issues with the service, suggesting that customer service interactions may serve as a red flag for potential churn.
3. **International Plan**: Customers without an international plan exhibit a significantly higher median churn rate. This suggests that the presence or absence of an international plan plays an important role in customer retention. The higher churn rate among customers without an international plan highlights the importance of this feature in predicting customer churn. Conversely, customers with an international plan appear to have a lower churn rate, indicating potential value in offering or maintaining such plans for better customer retention.

4. **VoiceMail Plan**: The analysis also suggests that customers who subscribe to a voicemail plan have a lower churn rate. This indicates that the voicemail service might add significant value to customers, encouraging them to stay with the company.

Feature Engineering: Feature importance analysis was conducted to identify the key features influencing churn. The results revealed that "Total day charge" and "Total day minutes" are the most influential features in predicting churn. Based on these findings, additional feature engineering was performed. Right-skewed and left-skewed features were treated differently in the preprocessing pipeline, with right-skewed features being transformed using a log transformation and left-skewed features being handled using a reciprocal log transformation.

Modeling and Metrics: For the modeling phase, we chose the following algorithms based on the size and complexity of the dataset:

- **Logistic Regression** and **Decision Tree**: Simple models that provide an initial baseline due to their interpretability and ease of use.
- **Random Forest and Gradient Boosting**: A more robust and accurate model, which offers higher performance.

To evaluate the models, we used the following metrics:

- **F1-Score**
- **Recall**
- **Precision**

Key Findings from Modeling:

- **Random Forest**: The Random Forest model exhibited signs of overfitting, with much better performance on the training data than on the test data. This discrepancy suggests that the model may not generalize well to unseen data.
- **Gradient Boosting**: Gradient Boosting demonstrated more consistent performance between training and testing data, suggesting that it is more robust and less prone to overfitting. It showed marginally better Precision and F1-Score compared to the other models, and the metrics for training and testing were closer in value, indicating better generalization.

Given these findings, **Gradient Boosting** appears to be the superior model for this task. Its balanced performance on both training and test data, along with slightly better Precision and F1-Score, makes it the most reliable model for predicting customer churn in this dataset.

Conclusion: Based on the analysis, the most important factors in predicting customer churn are the presence of an international plan, the frequency of customer service interactions, and the total charges and minutes related to different time periods

Notebook link