Title : An Analysis of Customer Churn Prediction Using Classification Models

Author : Pratik Ganguli (pgan501)

The ***CellPhoneChurn*** dataset captures customer usage behaviour and service interactions across 2,151 records to identify churn patterns and improve retention for a mobile service provider. This analysis examines the relationship between ***CustServCalls***, a discrete independent variable representing the number of customer service calls made, ***MonthlyCharge***, a continuous independent variable indicating the average monthly bill, and ***DayMins***, a continuous independent variable measuring the average daytime minutes used per month, as these factors often influence customer dissatisfaction and churn. A ***XGBoost Classification Model*** is applied to predict ***Churn***, a binary dependent variable indicating whether a customer cancelled their service. These insights will help the company proactively address potential churn risks, improving customer satisfaction and reducing revenue loss.

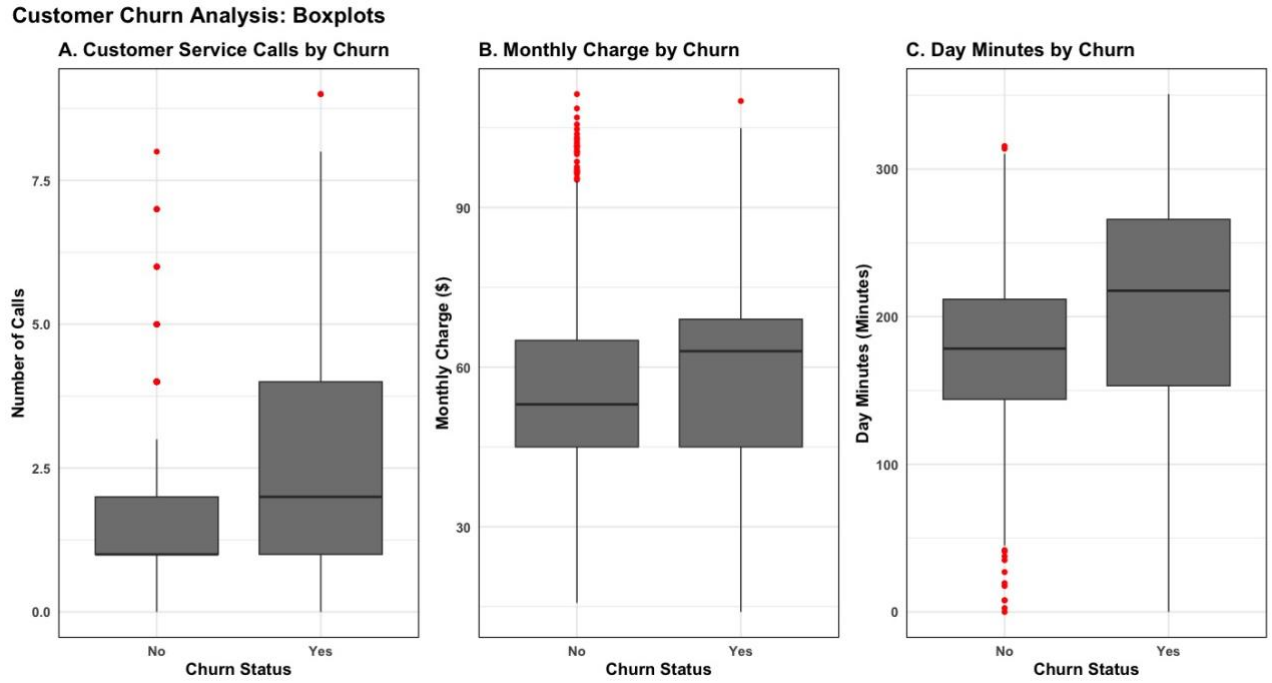

Figure 1.1 *Customer Churn Analysis: Key Metrics Comparison*

	CustServCalls		MonthlyCharge		DayMins	
Churn	**YES**	**NO**	**YES**	**NO**	**YES**	**NO**
Min.	0.00	0.000	14.00	15.70	0.0	0.0
1st Qu.	1.00	1.000	45.00	45.00	153.2	144.0
Median	2.00	1.000	63.00	53.00	217.6	178.4
Mean	2.23	1.479	59.19	56.23	206.9	176.6
3rd Qu.	4.00	2.000	69.00	65.03	265.9	211.8
Max.	9.00	8.000	110.00	111.30	350.8	315.6

Table 1.1 *Summary Statistics of Customer Service Calls (Frequency), Monthly Charges ($), and Daytime Minutes (Minutes) by Churn Status*

Churn	
YES	**NO**
22.45 %	77.54 %

Table 1.2 *Distribution of Customer Churn Proportions*

The summary statistics[Table 1.1] highlight key differences between customers who churned and those who remained. The overall churn rate is 22.45%[Table 1.2], with churned customers exhibiting a higher median MonthlyCharge($63) compared to non-churned customers($53), suggesting that increased billing may contribute to customer attrition. Additionally, the DayMins variable indicates that churned customers have a higher median usage(217.6 minutes) than those who did not churn(178.4 minutes), potentially linking higher daytime usage to an increased likelihood of churn.

The boxplots in[Figure 1.1] provide further insights into these relationships. [Figure 1.1.A]shows that churned customers tend to make a higher number of customer service calls, suggesting dissatisfaction or unresolved issues. [Figure 1.1.B] and [Figure 1.1.C] illustrate that churned customers tend to have higher monthly charges and greater daytime usage, aligning with the pattern observed in the summary statistics and highlighting potential cost-related drivers of churn. These findings emphasize the importance of targeted retention strategies, particularly for high-usage and high-billing customers, to mitigate churn and improve customer satisfaction.

Metric	Accuracy	AUC (Area Under Curve)	Sensitivity	Specificity	Precision
Training Data (Average)	0.856 ≈ 85.6%	0.805	0.913 ≈ 91.3%	0.657 ≈ 65.7%	0.902 ≈ 90.2%
Testing Data	0.847 ≈ 84.7%	0.774	0.916 ≈ 91.6%	0.607 ≈ 60.7%	0.890 ≈ 89%

Table 1.3 *XGBoost Classification Matrix for Training and Testing Data*

The dataset was partitioned using a stratified 70%-30% train-test split, and **SMOTE**(Synthetic Minority Over-sampling Technique) was applied to mitigate class imbalance, ensuring a representative distribution of churned and non-churned customers. Additionally, key preprocessing steps were implemented, including normalizing numeric predictors, to enhance model performance. Among the machine learning classification models tested, ***XGBoost*** demonstrated the highest **sensitivity** across multiple test runs, achieving 91.3% in training and 91.6% in testing, indicating the model's strong ability to correctly identify over 91%[Table 1.3] of actual churners, making it highly effective for churn prediction.

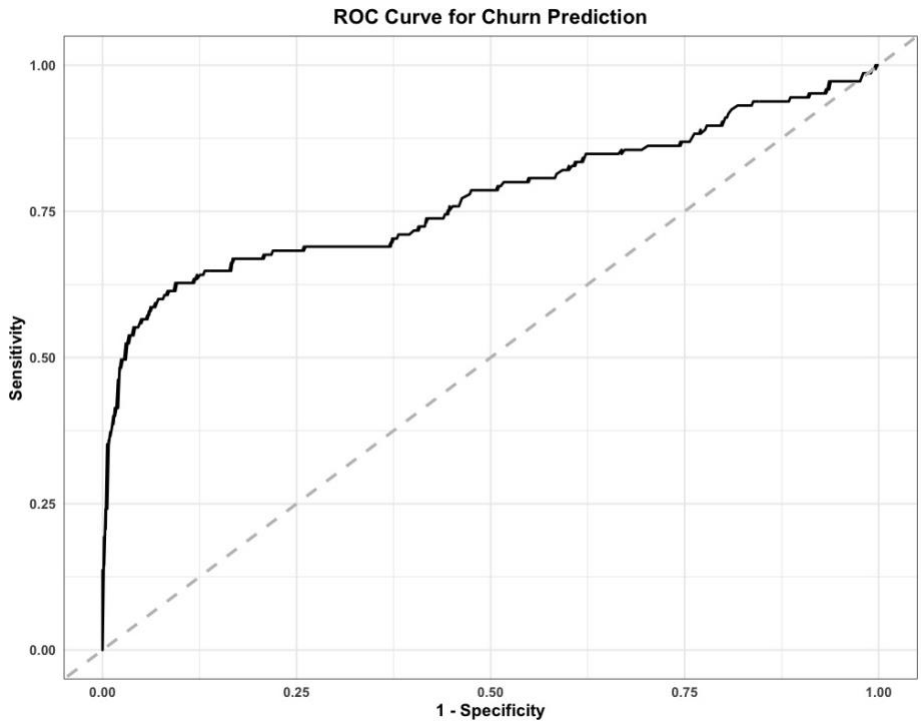

Figure 1.2 *Model Performance Assessment Using ROC Curve (Testing Data)*

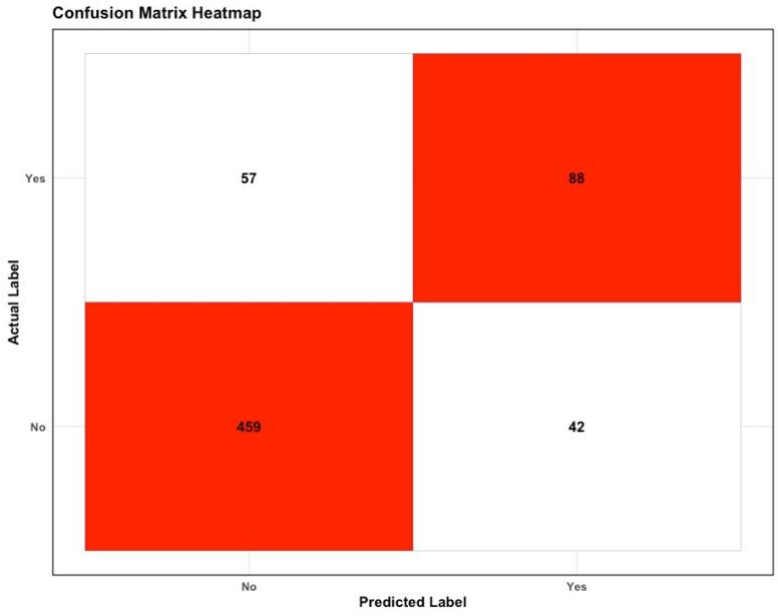

Figure 1.3 *Visual Representation of the Confusion Matrix (Testing Data)*

The model's **ROC-AUC** scores of 0.805(training) and 0.774(testing)[Figure 1.2 & Table 1.3] highlight its effectiveness in distinguishing churners from non-churners. The confusion matrix[Figure 1.3] shows that the model correctly predicted 88 churned customers(True Positives) but misclassified 57 as non-churners(False Negatives), which could lead to missed retention opportunities. Additionally, 42 non-churners were incorrectly classified as churners(False Positives), leading to unnecessary retention efforts. The model achieved an **accuracy** of 84.7%, with a precision of 89%, a **specificity** of 60.7%, and a strong **sensitivity** of 91.6%[Table 1.2], reinforcing its reliability in detecting at-risk customers. These insights are crucial for enhancing retention strategies and enabling proactive engagement with high-risk customers.

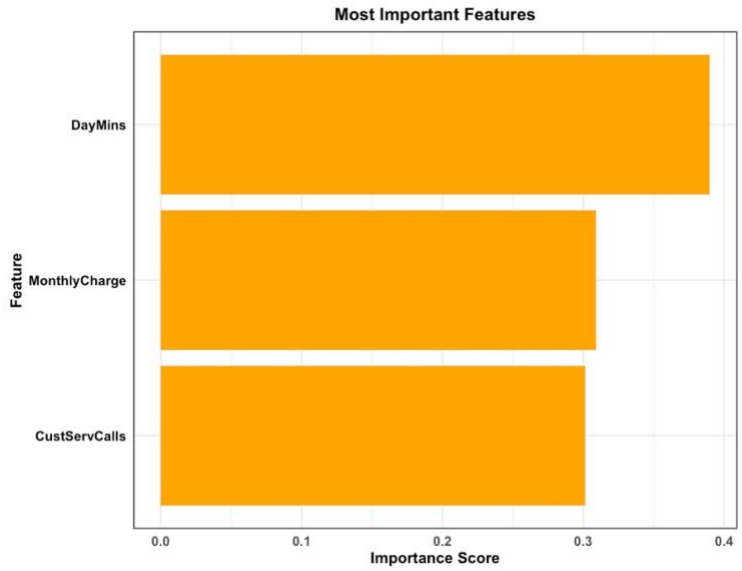

Figure 1.4 *Feature Importance Scores in Customer Churn Prediction Using XGBoost (Testing Data)*

In summary, this analysis highlights key factors influencing customer churn, with DayMins emerging as the strongest predictor, followed by MonthlyCharge and CustServCalls, as shown in the feature importance plot[Figure 1.4], indicating that higher usage, elevated billing, and increased service interactions contribute significantly to churn risk. The XGBoost model, with its high sensitivity(91.6%), effectively identifies at-risk customers, enabling targeted retention efforts. The model's strong predictive performance ensures that the company can proactively address churn risks, enhance customer satisfaction, and implement data-driven strategies to improve long-term revenue stability.