DSC1107

Predicting Customer Churn

A paper by Cuerdo and Percia

TOPIC OUTLINE

Introduction

Methodology

Results and Discussion

Conclusion and Recommendations

Introduction

CUSTOMER CHURN

One way of measuring a key performance in the telecom industry is providing customer churn. Identifying customers who are at risk of leaving allows companies to take action in retention measures. Thus. This project uses the Orange Telecom churn dataset to build and evaluate predictive models that classify whether a customer is likely to churn. With a variety of modeling approaches, this study compares performance in terms of accuracy, interpretability, and business applicability.

Methodology

DATASET

The dataset consists of two files: churn-bigml-80.csv for training and cross-validation, and churn-bigml-20.csv for final model evaluation. Each record contains customer usage patterns, plans, and a binary Churn label (1 = churned, 0 = stayed).

Methodology

PRE-PROCESSING STEPS

From the dataset, the researchers removed non-predictive features like State and Area.code. Then variables like Churn are converted from text to binary (0/1). Categorical variables (International.plan, Voice.mail.plan) are also converted using one-hot encoding.

Methodology

MODELS APPLIED

Models were split into two groups:

- Regression-Based Models: Logistic Regression, Ridge Regression, Lasso Regression.

Tree-Based Models: Decision Tree, Random Forest, XGBoost (Gradient Boosting).

Methodology

EVALUATION METRICS

- Accuracy
- Precision, Recall, F1-score
- AUC (Area Under the ROC Curve)
- Feature Importance (for tree-based models and Lasso)

Results & Discussion

Churn	n	Percentage
0	2278	85.44636
1	388	14.55364

Table 1 Customer Churn Distribution

Table 1 shows the breakdown of customer churn in the training dataset. From the table, a total of 2278 customers did not churn, which means that they stayed with the company. Meanwhile, 388 customers (or 14.55%) did churn - they left or cancelled the service.

Results & Discussion

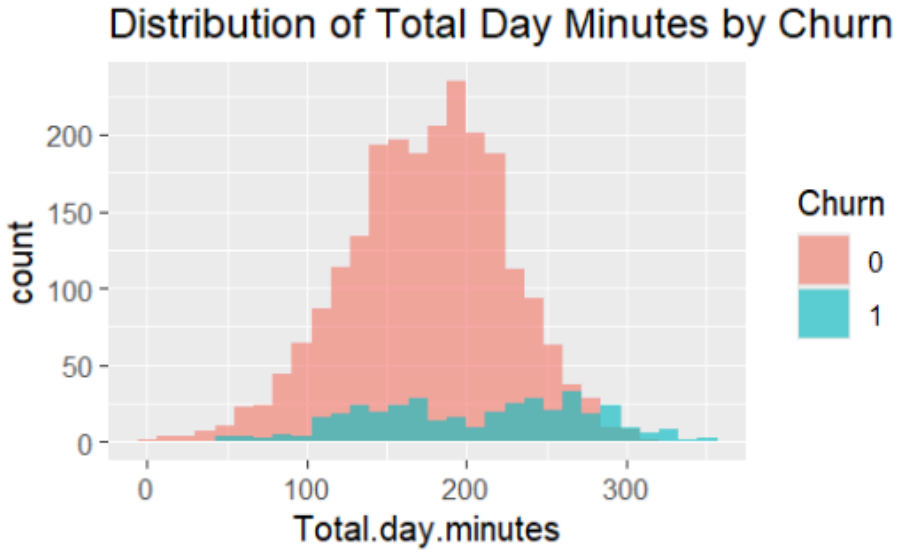

Figure 1. Histogram Distribution of Total Day Minutes by Churn.

Figure 1 shows the distribution of total day minutes used by customers, segmented whether they churned or not. Most non-churning customers (red) used between 150 to 250 minutes per day. This is where the red bars are tallest. Meanwhile Churning customers (blue) are more spread out and more heavily represented at the higher end of day minutes (e.g., 250–350 minutes). This suggests that customers who churned tended to use more day minutes on average compared to those who stayed.

Results & Discussion

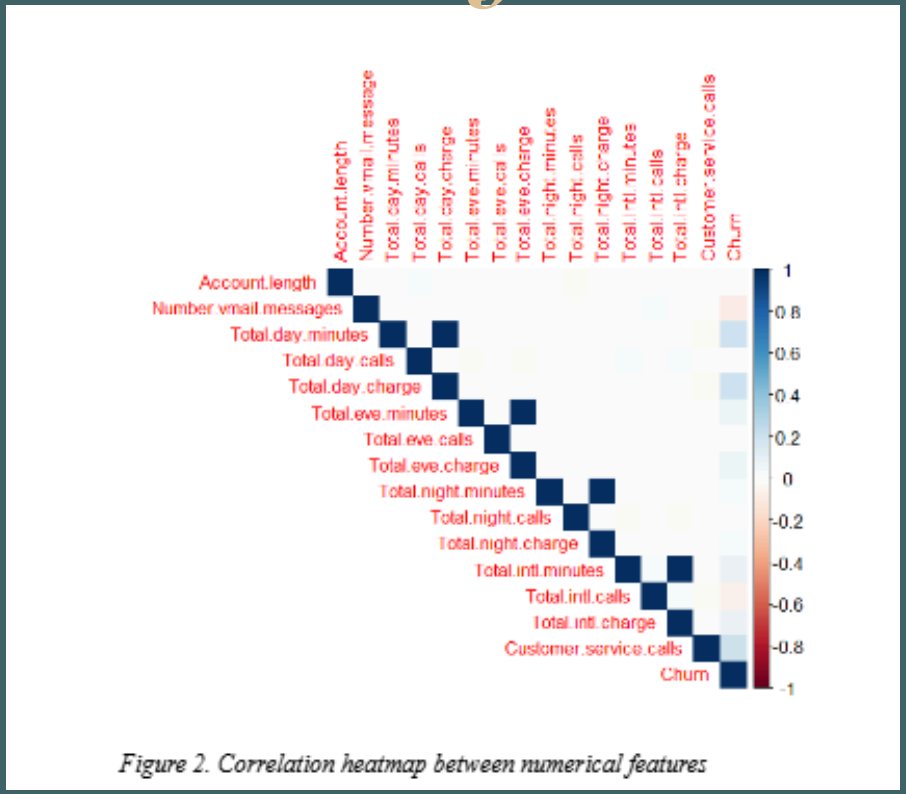

Figure 2. Correlation heatmap between numerical features

Figure 2 shows the correlation heatmap between the numerical features and its variables. From the figure above, it shows that Total.day.minutes and Total.day.charge have a very high positive correlation. Customer.service.calls have a moderative positive relation with Churn, suggesting that day-time users are more likely to leave. Total.night.minutes or Total.intl. calls show weak positive correlation.

Results & Discussion

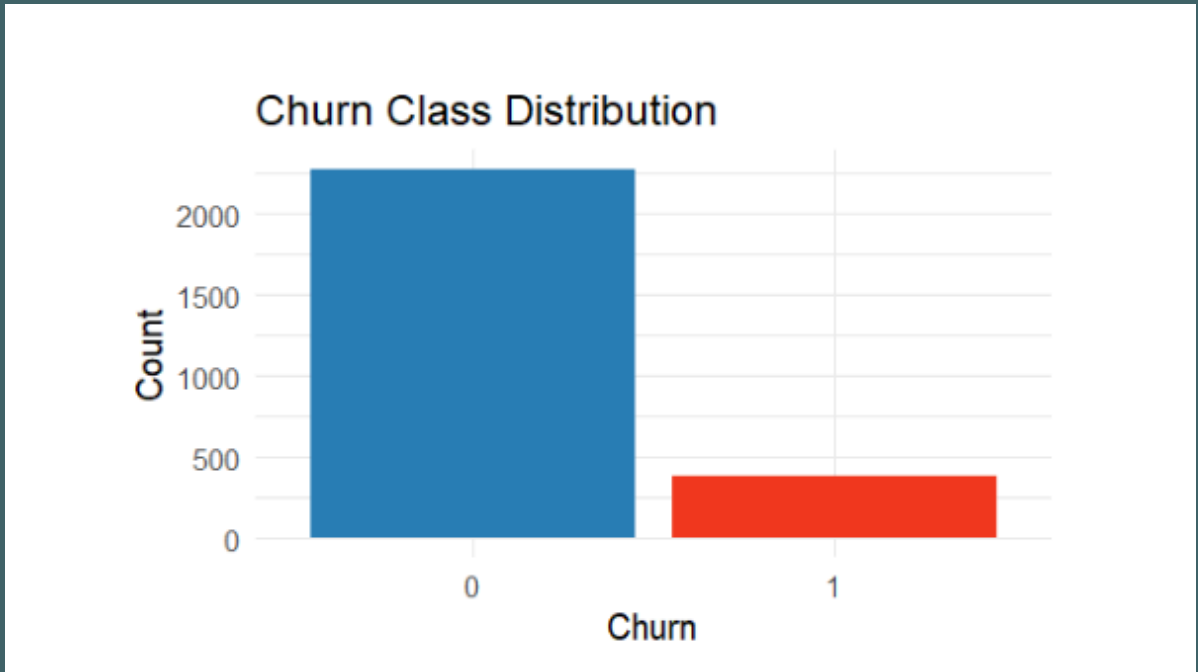

Figure 3. Churn Class Distribution

Figure 3 above shows the churn class distribution that aims to check class imbalance. From the figure above, The distribution above confirms the Churn rate, which contains that around 2000 are customers who did not churn, and those who did are only at around 300.

Results & Discussion

Logistic Regression Model

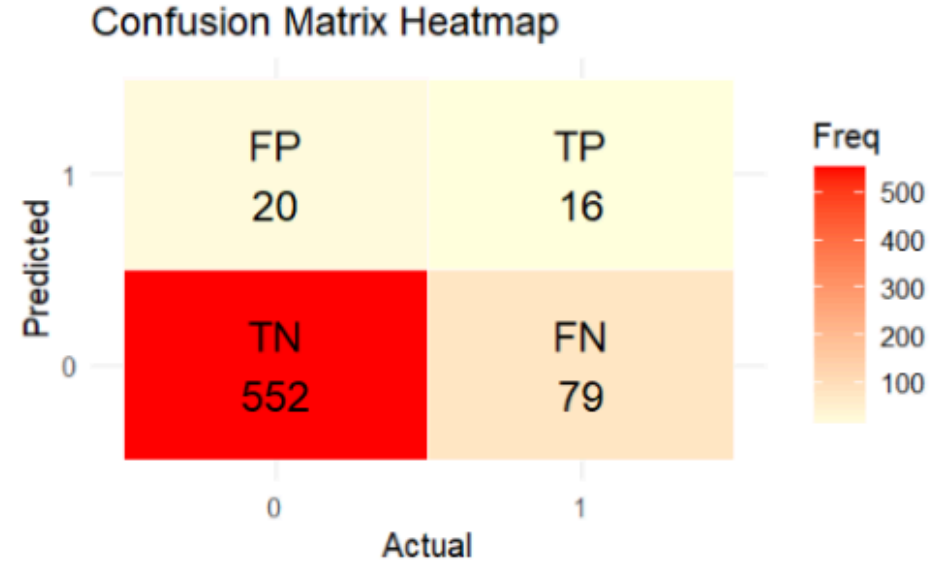

Figure 4. Confusion Matrix Heatmap for the Logistic Regression Model

Figure 4 shows a correlation heatmap for the confusion matrix of the logistic regression model. From the heatmap above, it shows that the model accurately predicted the true negatives, with 552 correctly identified cases. However, it struggled with identifying true positives, as only 16 were correctly classified out of 95 actual positive cases, indicating a low recall performance.

Results & Discussion

Logistic Regression Model

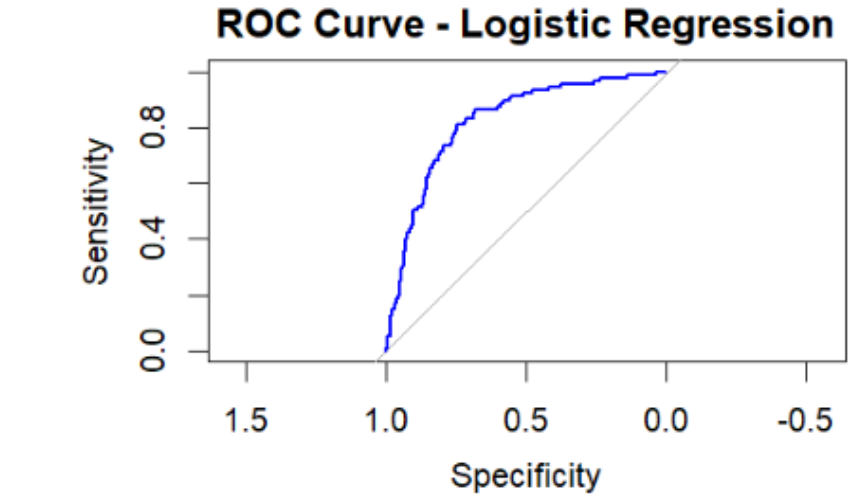

Figure 5. Receiving Operating Characteristic Curve for Logistic Regression Model

Figure 5 shows the ROC curve of the logistic regression model. From the plot above, the curve lies above the reference line, indicating a strong model performance. This suggests that the model is effective in distinguishing positive and negative classes. The steep rise of the curve shows a high true positive rate at low positive rates, further supporting the model's reliability.

Results & Discussion

Ridge and Lasso Model

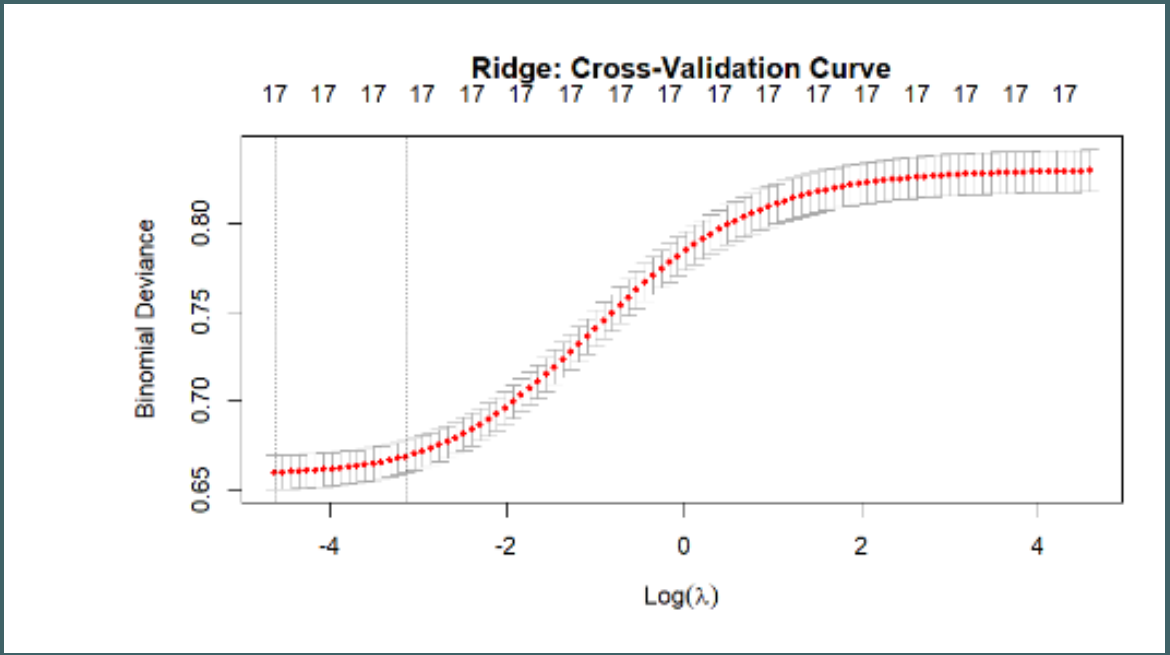

Figure 6 shows the cross-validation curve for the ridge logistic regression model. The model achieves its lowest binomial deviance at log(λ) ≈ -4, indicating this is the optimal level of regularization. The U-shaped curve highlights the trade-off between underfitting and overfitting: performance deteriorates when λ is too small (minimal regularization) or too large (excessive penalty). The chosen λ corresponds to a model that generalizes well without overfitting.

Results & Discussion

Ridge and Lasso Model

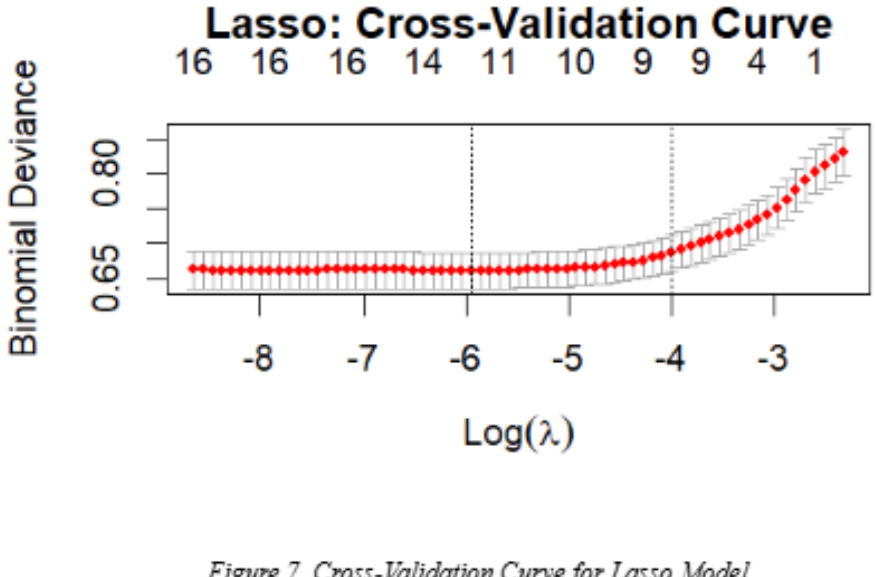

Figure 7. Cross-Validation Curve for Lasso Model

Figure 7 shows the cross-validation curve for the lasso logistic regression model. The model achieves its lowest binomial deviance at log(λ) ≈ -8, indicating that all 16 features were used but may overfit (train: 0.65, val: 0.80). The optimal λ maximizes validation performance while keeping enough predictive features.

Results & Discussion

Ridge and Lasso Model

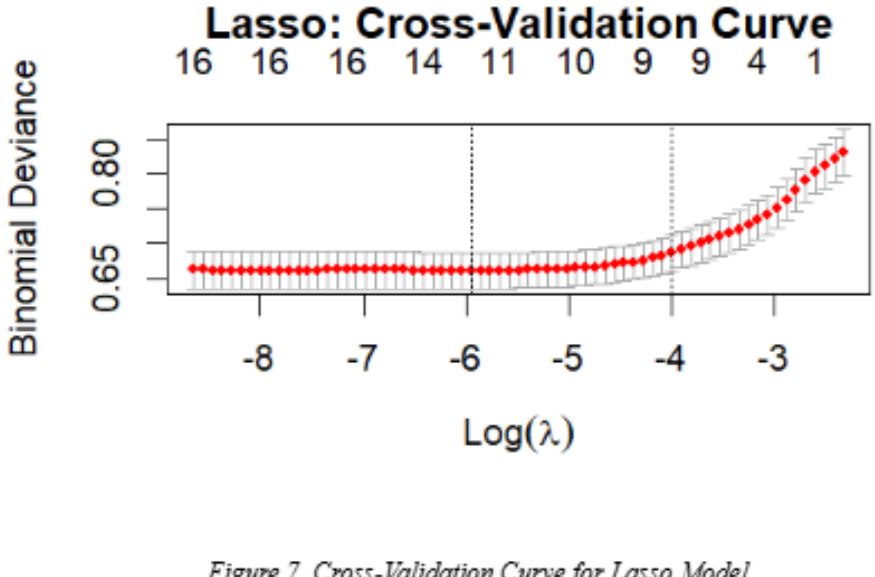

Figure 7. Cross-Validation Curve for Lasso Model

Figure 7 shows the cross-validation curve for the lasso logistic regression model. The model achieves its lowest binomial deviance at log(λ) ≈ -8, indicating that all 16 features were used but may overfit (train: 0.65, val: 0.80). The optimal λ maximizes validation performance while keeping enough predictive features.

Results & Discussion

Decision Tree

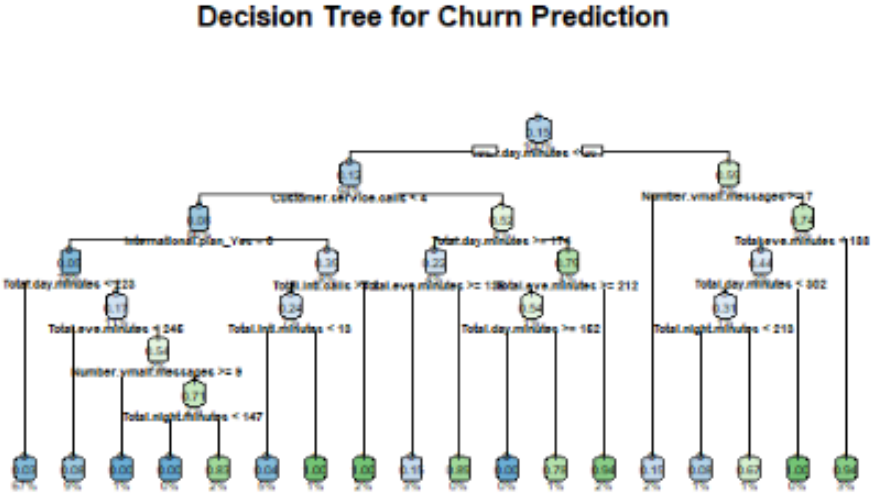

Figure 8. Decision Tree Model

Figure 8 shows the decision tree model for customer churn. From the plot above, it clearly shows that the decision tree handled the imbalances of the data, showing that its classification is balanced, has good sensitivity and specificity.

Results & Discussion

Random Forest Model

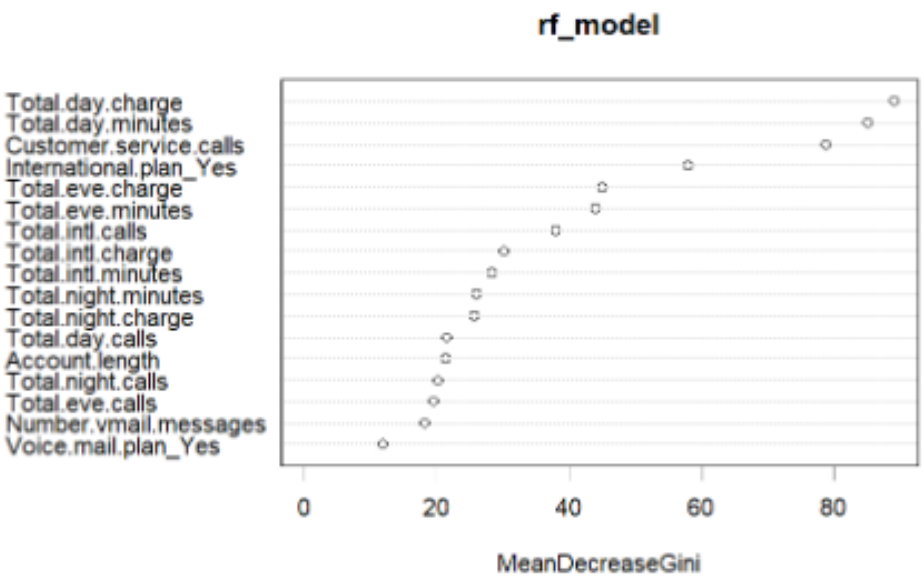

Figure 9. Feature Importance of Random Forest Model

Figure 9 shows the feature importance of the Random Forest Model. The plot reveals that Total day charge and Total day minutes are the most influential predictors, followed by Customer service calls and International plan_Yes. Call duration/charge metrics dominate the top contributors, while call frequency features (like Total day calls) and demographic factors (Account length) show lower importance.

Results & Discussion

Gradient Boosting Model

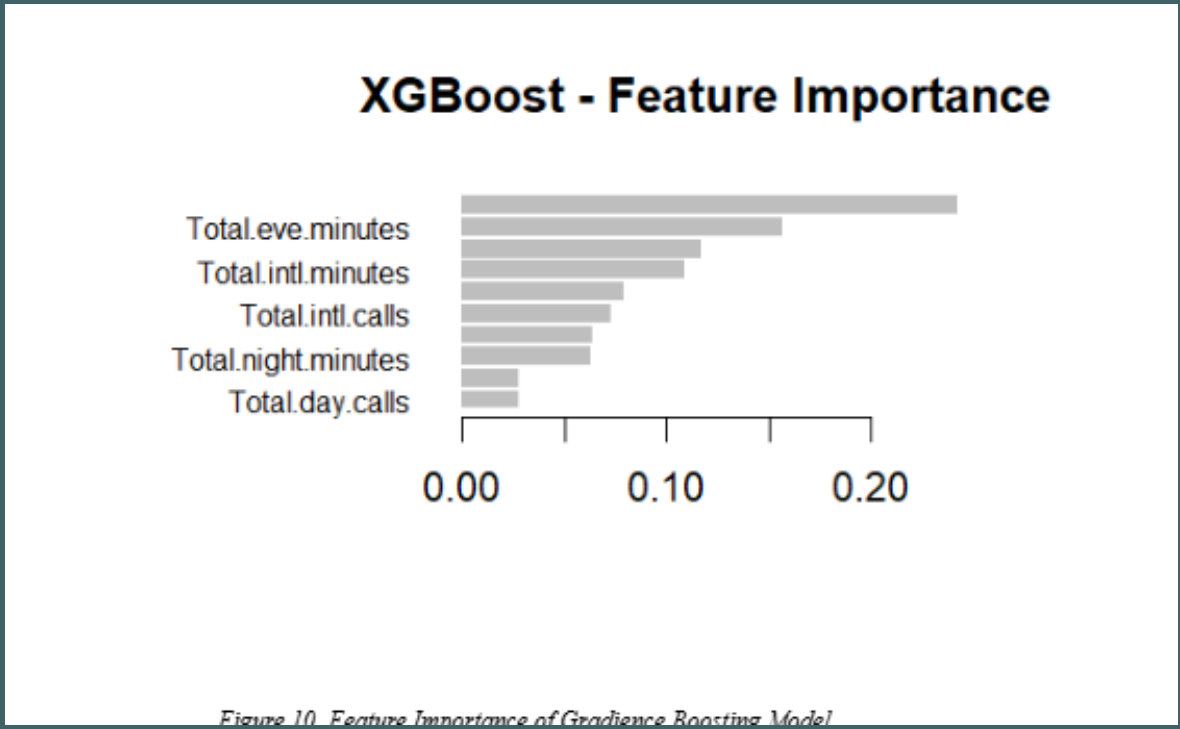

Figure 10. Feature Importance of Gradience Boosting Model

Figure 10 shows the feature importance of the gradience boosting model. From the plot above, it shows thatTotal.eye.minutes is the most influential predictor, followed by Total.intl.minutes and Total.night.minutes, while call-related features (Total.intl.calls, Total.day.calls) contribute less. This suggests call duration metrics drive predictions more than call frequency. The scale (0.00–0.20) indicates normalized importance, with the top feature covering ~20% of the model's decision power.

Conclusion & Recommendation

Among the models Tested:

- XGBoost performed best in terms of AUC and accuracy.
- Random Forest offered a close second with high accuracy and reliable feature importance rankings.
- Logistic Regression remains a strong baseline due to its interpretability and decent performance.

With this, this implies that Orange Telecom can use these models to:

- Please contact at-risk customers.
- Customize retention strategies based on key churn predictors.
- Improve service by analyzing churn-driving behaviors like frequent service calls or high daytime usage.

THANK YOU