Predicting Customer Churn with Regression-Based and Tree-Based Methods

A research paper

In partial fulfillment in the requirements of DSC1107 Data Mining and Wrangling,

S.Y. 2024-2025

Cuerdo, Naomi Hannah A.

Percia, Kyte Daiter M.

May 2025

I. Introduction

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.

II. 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).

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.

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).

Evaluation Metrics

Models were evaluated on:

- Accuracy
- Precision, Recall, F1-score
- ROC AUC (Area Under the ROC Curve)

- Feature Importance (for tree-based models and Lasso)

Furthermore, Train-test split and cross-validation tests were used in appropriate methods.

III. Results & Discussion

A. Exploratory Data Analysis

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.

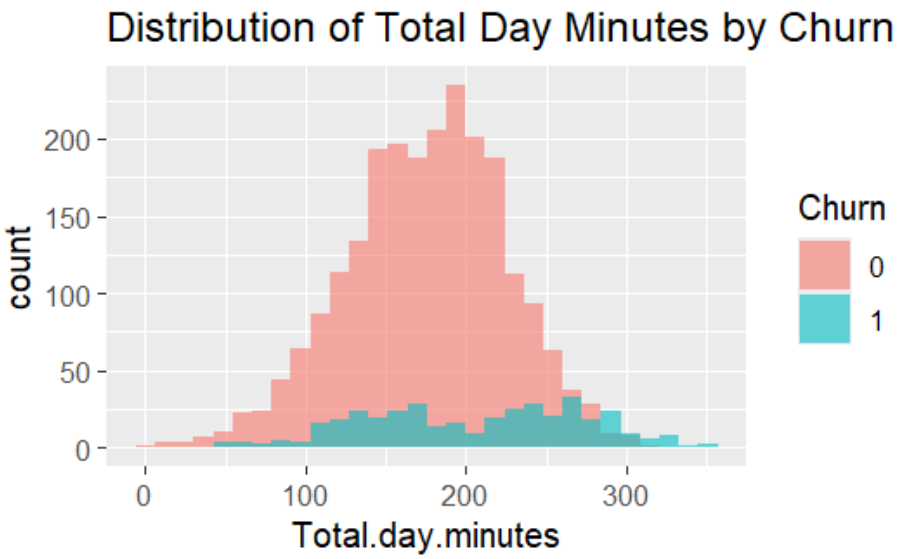

Figure 1. Histogram Distribution of Total Day Minutes by Churn.

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.

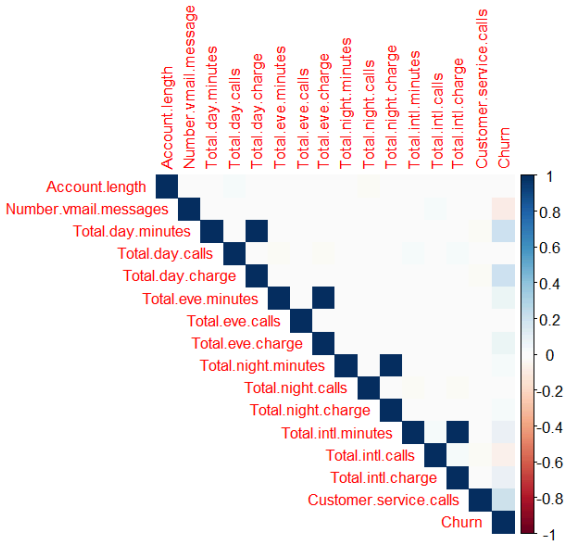

Figure 2. Correlation heatmap between numerical features

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.

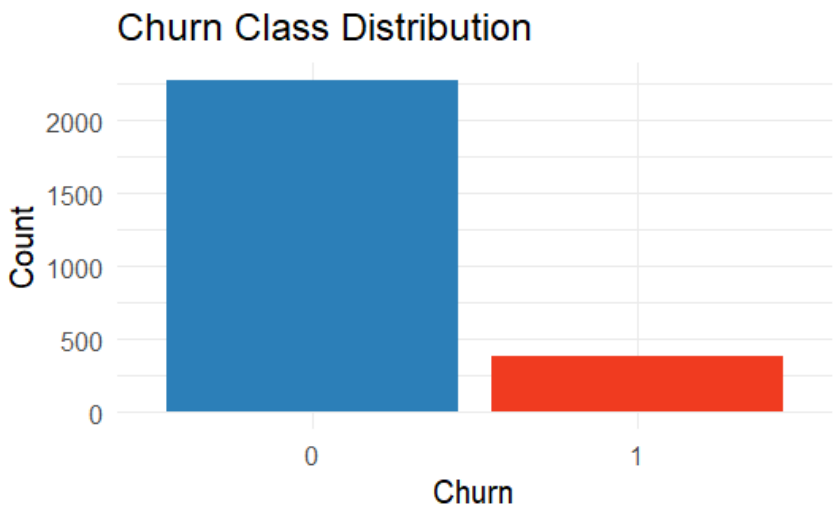

Figure 3. Churn Class Distribution

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.

B. Modeling and Comparison

This study will utilize models which are categorized into two groups: **Regression-based Models (i.e., Logistic Regression, Ridge and Lasso Regression),** and **Tree-based models (i.e., Decision Tree, Random Forest, Gradient Boosting).**

Logistic Regression Model

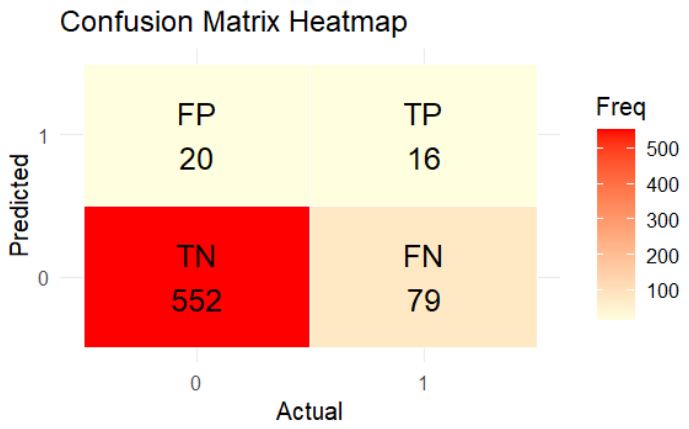

Figure 4. Confusion Matrix Heatmap for the Logistic Regression Model

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.

Metric	Accuracy	95% CI	Kappa	Mcnemat 's Test P-Value	Sensitivit y	Specificit y	Pos Pred Value	Neg Pred Value	Prevalenc e	Detection Rate	Detection Prevalenc e	Balanced Accuracy	Area under the Curve
Value	0.8516	(0.8223, 0.8777)	0.1801	0.000000 0569	0.965	0.1684	0.8748	0.4444	0.8576	0.8276	0.946	0.5667	0.826

Table 2. Confusion Matrix and Classification Metrics for the Logistic Regression Model

Table 2 summarizes the classification results of the logistic regression model. From the table, The model achieved an overall accuracy of 95.65%, with high sensitivity (98.43%) and a balanced accuracy of 88.69%. The Kappa statistic was 0.813, indicating strong agreement between predicted and actual classifications.

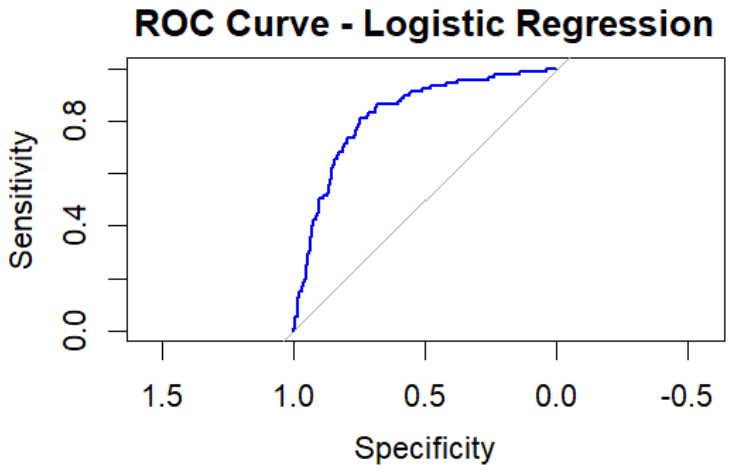

Figure 5. Receiving Operating Characteristic Curve for Logistic Regression Model

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.

Ridge and Lasso Model

Metric	Accuracy	95% CI	Kappa	Mcnemar's Test P-Value	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Prevalence	Detection Rate	Detection Prevalence	Balanced Accuracy	Area under the Curve
Value	0.8546	(0.8255, 0.8805)	0.1772	0.00000 0003071	0.9703	0.1579	0.874	0.4688	0.8576	0.8321	0.952	0.5641	0.8255

Table 3. Confusion Matrix and Classification Metrics for the Ridge Model

Table 3 shows the confusion matrix and classification metrics for the ridge model. From the table, the model achieved a high accuracy of 85.46%, largely due to the dominant presence of the negative class in the dataset. However, the model demonstrated poor sensitivity, with 80 false negatives compared to only 15 true positives. With an AUC of 0.8255, while it indicates a good discrimination ability, the current classification threshold fails to capitalize on this, resulting in

low specificity (15.79%) and a high false negative rate. While the ridge regularization likely improved coefficient stability and reduced overfitting, but did not resolve the issue of class imbalance.

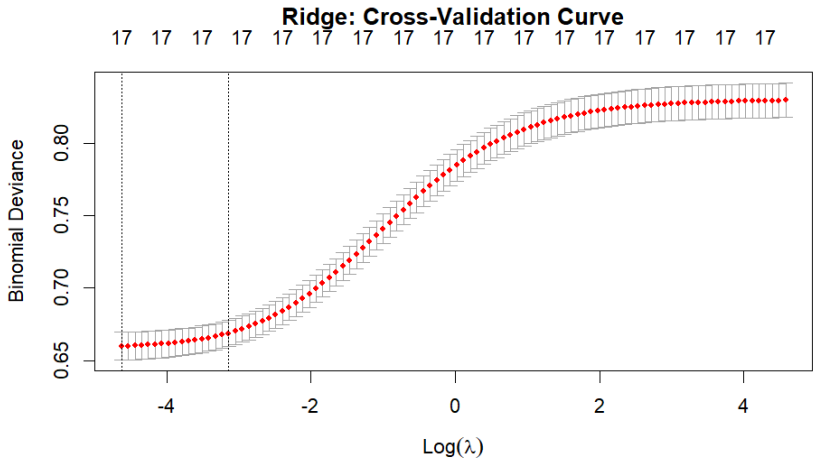

Figure 6. Cross - Validation Curve for Ridge Logistic Model

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.

Metric	Accuracy	95% CI	Kappa	Mcnemar's Test P-Value	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Prevalence	Detection Rate	Detection Prevalence	Balanced Accuracy	Area under the Curve
Value	0.8531	(0.8239, 0.8791)	0.1737	0.00000 0007186	0.9685	0.1579	0.8738	0.4545	0.8576	0.8306	0.9505	0.5632	0.825

Table 4. Confusion Matrix and Classification Metrics for the Lasso Model

Table 4 shows the confusion matrix and classification metrics for the lasso model. From the table, the model achieved a high accuracy of 82.5%, which means that the model performs well in detecting class 0. However, the model demonstrated poor specificity (0.16), and Kappa

(0.17%), which indicates that the model performed poorly with class 1, and the agreement with actual labels is weak. Overall, the model strongly favors class 0, likely due to class imbalance (85.76% prevalence).

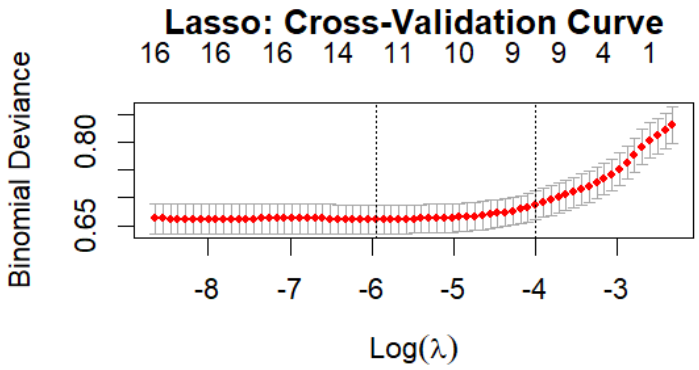

Figure 7. Cross-Validation Curve for Lasso Model

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.

Decision Tree

Metric	Accuracy	95% CI	Kappa	Mcnemar's Test P-Value	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Prevalence	Detection Rate	Detection Prevalence	Balanced Accuracy	Area under the Curve
Value	0.9565	(0.9382, 0.9707)	0.813	0.06332	0.9843	0.7895	0.9657	0.8929	0.8576	0.8441	0.8741	0.8869	0.825

Table 5. Confusion Matrix and Classification Metrics for the Decision Tree Model

Table 5 shows the confusion matrix and classification metrics for the decision tree. From the table, the model achieved a high accuracy of 95.65%, which indicates that the model handles the class imbalance much better.. However, the model demonstrated minor classification (9 false positives, 20 false negatives), but it is not that concerning,

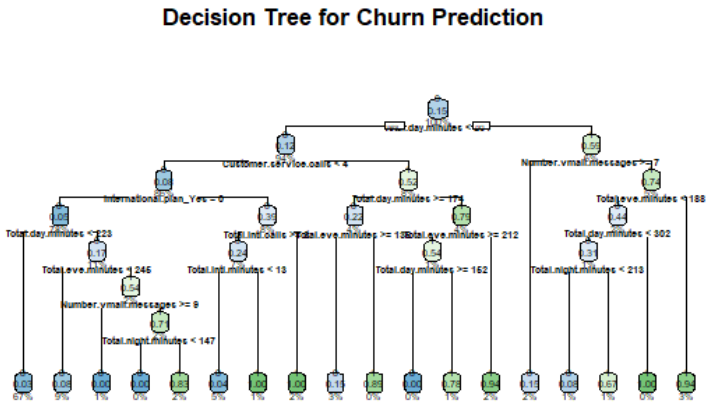

Figure 8. Decision Tree Model

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.

Random Forest

Metric	Accuracy	95% CI	Kappa	Mcnemar's Test P-Value	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Prevalence	Detection Rate	Detection Prevalence	Balanced Accuracy	Area under the Curve
Value	0.955	(0.9382, 0.9707)	0.7982	0.00052 26	0.9913	0.7368	0.9578	0.9333	0.8576	0.8501	0.8876	0.8641	N/A

Table 6. Confusion Matrix and Classification Metrics for the Random Forest Model

Table 6 shows the confusion matrix and classification metrics for the decision tree. From the table, the model performs almost as well as the decision tree model, with slightly better sensitivity and fewer false positives. Although the specificity and Kappa is slightly lower, the model still performs well.

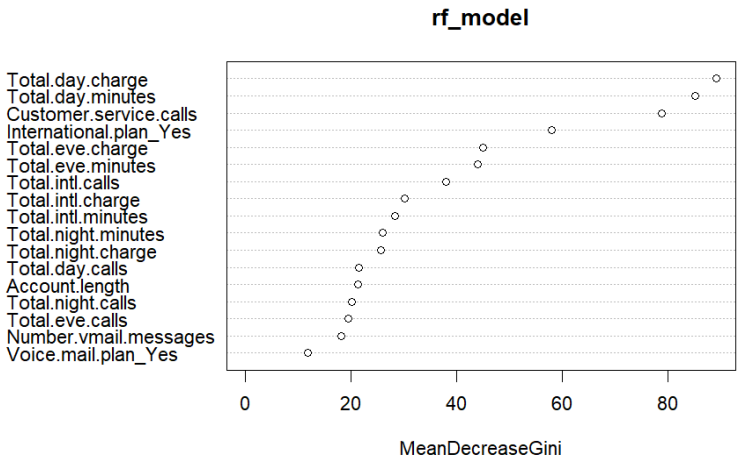

Figure 9. Feature Importance of Random Forest Model

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.

Gradient Boosting Model

Metric	Accuracy	95% CI	Kappa	Mcnemar's Test P-Value	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Prevalence	Detection Rate	Detection Prevalence	Balanced Accuracy	Area under the Curve
Value	0.958	(0.9399, 0.9719)	0.8169	0.01402	0.9878	0.7789	0.9642	0.9136	0.8576	0.8471	0.8786	0.8834	0.928

Table 7. Confusion Matrix and Classification Metrics for the Random Forest Model

Table 7 shows the confusion matrix and classification metrics for the gradient boosting model. From the table, the model achieved the highest performance among all models, with an accuracy of 95.8%, a curve of 0.928, and a strong Kappa score of 0.8169, indicating excellent

agreement between predictions and actual outcomes. The model also distributed high sensitivity (0.9878) and specificity (0.7789), showing effectiveness in handling imbalance data.

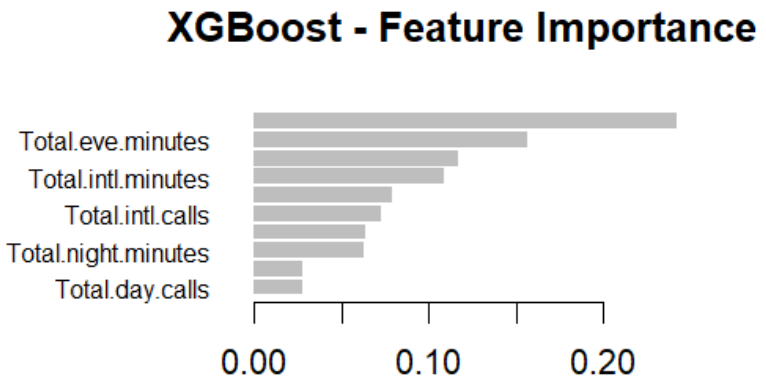

Figure 10. Feature Importance of Gradience Boosting Model

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.

IV. Conclusion & Recommendation

This study aims to determine which model can effectively classify churn behavior among telecom customers. 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:

- Proactively reach out to at-risk customers.
- Customize retention strategies based on key churn predictors.
- Improve service by analyzing churn-driving behaviors like frequent service calls or high day-time usage.

V. Appendix

a. Github link

https://github.com/aluminmi/DSC1107/blob/main/SA2/SA2_CUERDO_PERCIA.md