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

This section outlines the design, implementation, and evaluation of machine learning models to predict customer churn using the Orange Telecom Churn dataset. Both regression-based and tree-based methods are explored, and their performance is compared using standard evaluation metrics.

## 1. Data Cleaning

The methodological approach taken in this study ensures model fairness, comparability, and real-world relevance. The choice to drop the 'State' feature and encode categorical variables such as 'International plan' and 'Voice mail plan' stems from their potential to either bias the model with irrelevant location-specific data or cause algorithmic errors due to non-numeric representations. One-hot encoding maintains model flexibility by treating each category independently, which is critical for models like logistic regression that assume linearity in feature impact. The conversion of the target variable into binary form allows it to align with the expectations of classification algorithms.  
  
Initial data preparation involved dropping non-predictive identifiers such as phone numbers and state information. The churn column was converted into binary format (0 = No, 1 = Yes). Categorical variables were processed using one-hot encoding to make them suitable for machine learning models.

## 2. Exploratory Data Analysis (EDA)

EDA revealed a significant class imbalance, with far more customers staying than churning. This imbalance is critical as it influences the sensitivity of the models, particularly logistic regression—which may predict the majority class (non-churn) disproportionately, leading to misleadingly high accuracy but poor recall for the minority class (churners). The visualizations of call duration and plan usage were essential in highlighting potential behavioral signals. For instance, higher total day minutes may imply heavy reliance on the service, possibly reducing churn likelihood unless paired with high service dissatisfaction.

We examined overall churn rates, distributions of numerical features like call minutes and customer service calls and identified class imbalance. Correlation matrices were also visualized to detect potential multicollinearity.

## 3. Modeling and Evaluation

The selection of logistic regression and decision tree models provides a contrast between interpretability and non-linear learning capacity. Logistic regression offers clear coefficient-based reasoning, showing how each feature influences the probability of churn. Decision trees, in contrast, flexibly model interactions and splits in the feature space that might capture more nuanced customer behavior patterns. Their ability to segment the dataset into high-risk and low-risk churn groups is particularly advantageous in real-world deployment.  
  
We trained the following models: (1) Logistic Regression: A baseline interpretable model for binary classification. (2) Ridge/Lasso Regression: Used for regularization and feature selection. (3) Decision Tree: Provides a non-linear model with interpretability. (4) Random Forest: An ensemble tree-based method that improves generalization.  
(5) Gradient Boosting: A powerful ensemble method effective for imbalanced data.  
  
Each model was evaluated using accuracy, precision, recall, F1-score, and ROC AUC, with cross-validation for reliable results. Feature importance was extracted from tree-based models and coefficients from regression models for interpretability.

## 4. Results and Discussion

Churn Distribution

A blue rectangular bar graph

AI-generated content may be incorrect.

This bar plot shows a significant **class imbalance** in the dataset: the majority of customers did not churn (No (0)), while a smaller fraction did (Yes (1)). This imbalance can lead to biased models that favor the majority class if not properly addressed. It highlights the need for performance metrics like **recall** and **F1-score**—not just accuracy—and possibly the use of techniques such as **resampling** (e.g., SMOTE) or **class-weighted models** to balance predictive focus.

Distribution of Call Minutes

A group of blue and black graphs

AI-generated content may be incorrect.

This set of histograms illustrates the distribution of usage patterns for total minutes spent on calls during the day, evening, night, and international time periods.

* **Day, Eve, and Night minutes** are roughly **normally distributed**, centered between 150 and 250 minutes, suggesting consistent service usage across the majority of customers.
* **International minutes**, in contrast, show a **tighter and lower range**, peaking around 10 minutes. This reflects the rarity and limited use of international calling.

These distributions are critical for model training because they indicate the scale and spread of key continuous variables, as well as help identify outliers that may disproportionately affect model performance.

Correlation Heatmap

A graph with red and blue squares

AI-generated content may be incorrect.

This heatmap reveals correlations among numerical features and their relationship with the churn outcome:

* **High correlation within charge-minute pairs** (e.g., total day minutes and total day charge) is expected, as charges are computed from usage.
* **Customer service calls** and **international plan usage** show modest positive correlations with **churn**, indicating that customers who frequently call customer support or have international plans are slightly more likely to leave.
* **Voice mail plan** has a weak negative correlation with churn, possibly suggesting that customers who opt into voicemail services are more engaged or satisfied.

While no single feature is highly correlated with churn, these moderate associations help guide feature importance analysis in model interpretation, especially for decision trees and logistic regression.

Modelling and Comparison (Regression-Based and Tree-Based)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| model | accuracy | precision | recall | f1 | roc\_auc |
|  |  |  |  |  |  |
| Logistic | 0.769 | 0.355 | 0.758 | 0.483 | 0.831 |
| Ridge (L2) | 0.769 | 0.355 | 0.758 | 0.483 | 0.831 |
| Lasso (L1) | 0.772 | 0.358 | 0.758 | 0.486 | 0.830 |
| Decision Tree | 0.921 | 0.719 | 0.726 | 0.723 | 0.840 |
| Random Forest | 0.949 | 0.969 | 0.663 | 0.788 | 0.919 |
| Gradient Boosting | 0.951 | 0.931 | 0.705 | 0.802 | 0.925 |

**The three regression-based models (Logistic, Ridge, Lasso) show nearly identical performance, with:**

* **Accuracy ~0.77**
* **High recall (~0.758)** but **very low precision (~0.355–0.358)**
* **Moderate F1-scores (~0.48)**

These models effectively **capture most churners** (high recall), but they **also misclassify many non-churners as churners** (low precision), making them prone to **false positives**. This behavior is expected in **linear models** when the boundary between classes is non-linear or highly complex. Regularization via L1 and L2 does not substantially alter the performance, but Lasso slightly improves precision and F1.

**These models are interpretable and useful for identifying *risk factors*, but their predictive power is limited due to linearity. In practice, they are suitable for flagging at-risk users early on but would require downstream filtering or business rules to avoid excessive false alarms.**

**Decision Tree’s accuracy jumps to 0.921**, significantly higher than linear models

* Balanced **precision (0.719)** and **recall (0.726)**
* **F1-score (0.723)** shows good harmonic performance
* Slight improvement in **ROC AUC (0.840)**

The decision tree captures **non-linear interactions** and **combinatorial patterns** missed by regression models. Its **transparent rule-based logic** makes it interpretable, although it may still **overfit** if not pruned carefully.

**Random Forest’s accuracy increases to 0.949**

* **Precision is exceptionally high (0.969)**, meaning almost every predicted churner *did* churn
* **Recall drops slightly (0.663)**, missing some churners
* **F1-score of 0.788**, striking a strong balance
* **ROC AUC = 0.919**, signaling excellent discriminative ability

Random Forests reduce overfitting by averaging across many trees, stabilizing performance and boosting generalization. It sacrifices some recall for higher **precision and accuracy**—ideal for minimizing **false positives**, which can reduce wasted retention efforts.

This model is strong when resources are limited and **only the most likely churners** should be targeted (high precision). However, the drop in recall means some churners may go undetected. It's excellent for confident action but not exhaustive detection.

**Gradient Boosting provide best accuracy (0.951)** and **best F1-score (0.802)**

* **Precision (0.931)** remains very high
* **Recall improves (0.705)** over Random Forest
* **ROC AUC = 0.925**, highest among all models

Gradient Boosting optimizes iteratively to correct errors, making it especially powerful in **imbalanced classification tasks**. It provides the best **overall balance** between **capturing churners** and **minimizing false positives**, as reflected in its top F1 and AUC scores.

Gradient Boosting is the **best performing model** among most metrics. It is particularly suited to real-world churn prediction where **both recall and precision are valuable**. It captures nuanced behavioral patterns while maintaining high confidence in its predictions.

**Top 10 Most Important Features – Random Forest**

|  |  |  |
| --- | --- | --- |
| Rank | Feature | Importance Score |
| 1 | Customer service calls | 0.1444 |
| 2 | Total day charge | 0.1255 |
| 3 | Total day minutes | 0.1245 |
| 4 | International plan (Yes) | 0.0928 |
| 5 | Total evening minutes | 0.0589 |
| 6 | Total evening charge | 0.0529 |
| 7 | Total international minutes | 0.0437 |
| 8 | Total international charge | 0.0427 |
| 9 | Total international calls | 0.0423 |
| 10 | Total night charge | 0.0415 |
|  |  |  |

The Random Forest model identified **Customer Service Calls** as the most important predictor of churn, suggesting that customers with frequent service issues are highly likely to leave. This is followed closely by **daytime usage metrics** (minutes and charges), which may reflect customer activity and plan suitability.

## 5. Conclusion and Interpretation

Among all models, Gradient Boosting yielded the highest ROC AUC and F1-score, suggesting strong predictive performance, especially under class imbalance. Logistic Regression was more interpretable but suffered from lower recall.  
  
Tree-based models demonstrated a better bias-variance trade-off, capturing complex patterns in the data. However, interpretability decreased with model complexity. In real-world telecom applications, our results suggest that a hybrid approach — using interpretable models for monitoring and complex models for prediction — may be optimal.  
  
The model can be used by telecom firms to proactively retain at-risk customers, improve service quality targeting, and optimize marketing outreach.

From a business perspective, the ability to accurately identify churn-prone customers has direct cost-saving implications. Targeted retention campaigns can be designed around the top predictive features—such as customer service calls or international plan usage—revealed through feature importance. For example, if high customer service calls correlate with churn, telecom providers can proactively intervene by improving resolution quality. Moreover, the model’s deployment should prioritize recall to minimize missed opportunities in preventing customer loss, even at the cost of some false positives.

## 6. Code Appendix

<https://github.com/Monferium/DSC1107>