2024

Predictive Customer Churn Analysis

DURGA DEEPAK VALLURI

1. Introduction

Customer churn is a critical concern in the telecom industry, directly impacting revenue and profitability. High churn rates often indicate customer dissatisfaction, intense competition, or ineffective engagement strategies.

Objective

The primary goal of this project is to predict customer churn using advanced data mining and statistical modeling techniques. By identifying high-risk customers, the company can proactively implement retention strategies and minimize revenue loss. The project spans several phases:

1. Data exploration and preprocessing.
2. Feature selection and engineering.
3. Model building and evaluation.
4. Deployment and actionable business recommendations.

---

2. Data Exploration and Preprocessing

Dataset Overview

- **Source:** Provided churn dataset.
- **Size:** 7,011 rows × 21 columns.
- **Key Attributes:**
    - **Demographics:** Gender, SeniorCitizen, Partner, Dependents.
    - **Account Information:** Tenure, Contract, PaymentMethod, PaperlessBilling.
    - **Service Features:** InternetService, OnlineSecurity, TechSupport, StreamingTV.
    - **Target Variable:** Churn, indicating whether a customer left (Yes) or stayed (No).

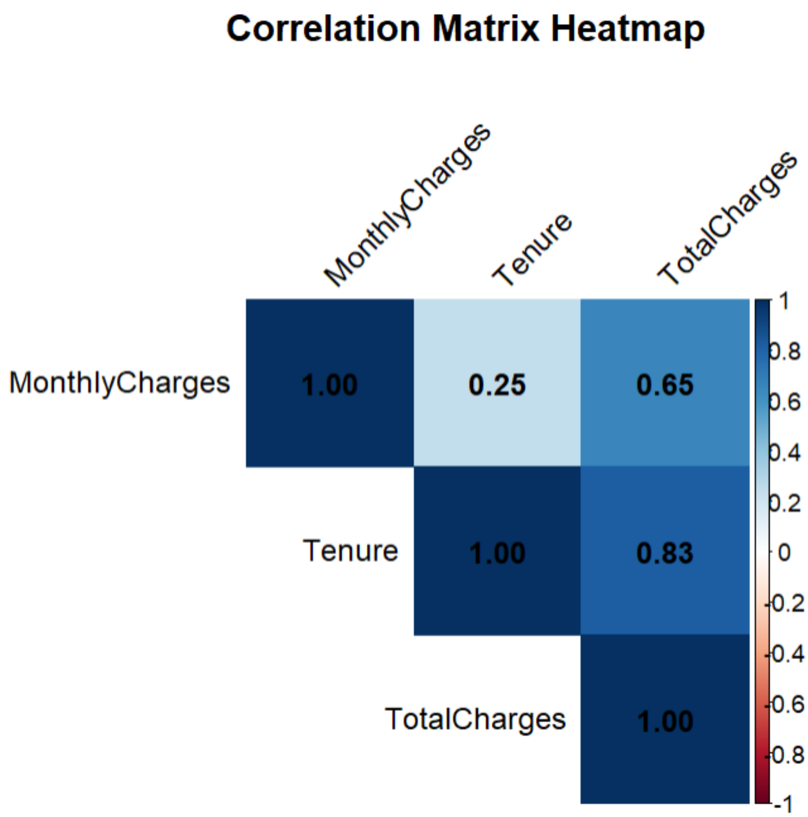

FIGURE 1: THE CORRELATION MATRIX HIGHLIGHTS RELATIONSHIPS BETWEEN NUMERICAL VARIABLES

- **Tenure and TotalCharges** have a strong positive correlation (r ≈ 0.83). This indicates that longer-tenured customers contribute more to total charges, reflecting sustained engagement.
- **MonthlyCharges and TotalCharges** show a moderate correlation (r ≈ 0.65), meaning monthly charges play a role in total revenue but are less critical than tenure.
- **Tenure and MonthlyCharges** have a weak correlation (r ≈ 0.25), suggesting that tenure does not directly influence the cost customers pay each month.

Takeaway: TotalCharges and Tenure are closely connected, making tenure a key focus for understanding churn and customer value.

Steps Undertaken

2.1 Data Cleaning

1. Removed customerID, as it does not provide predictive value.
2. Handled missing values using na.omit(), retaining valid rows for analysis.

2.2 Feature Transformation

- Encoded categorical variables as factors for model compatibility.
- Ensured Churn was a binary factor with valid levels (Yes and No).

2.3 Feature Engineering

To enrich the dataset:

- **Customer Lifetime Value (CLV):** Estimated as Tenure × MonthlyCharges.
- **Usage Pattern Count:** Number of "Yes" responses across service-related columns.

2.4 Exploratory Data Analysis

1. **Correlation Matrix**:

- **Tenure** and **TotalCharges** show a strong correlation ($r \approx 0.83$), highlighting the importance of customer longevity.
- Weak correlation between **MonthlyCharges** and **Tenure** suggests distinct impacts on churn.
- **Visualizations**:
  Bar Plots: High churn among month-to-month contracts and electronic check users.

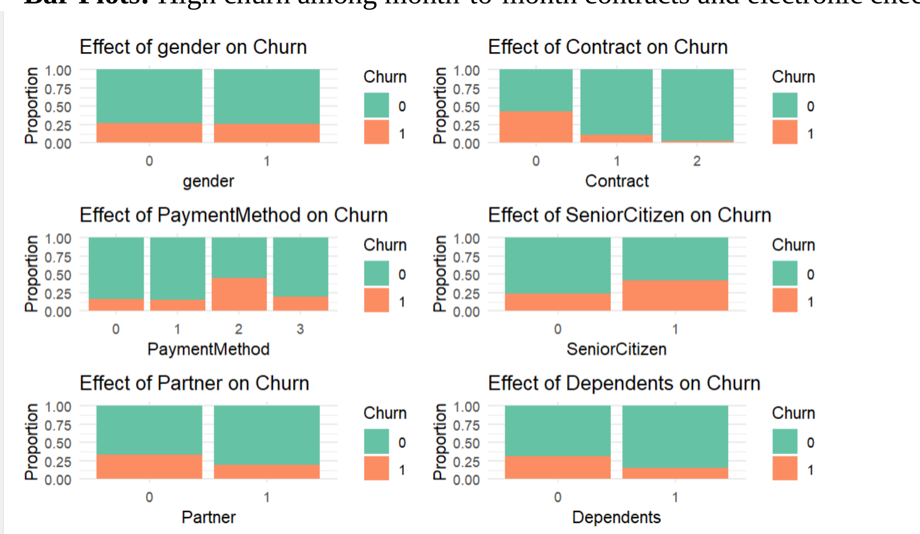

FIGURE 2: BAR PLOTS REVEAL HOW DIFFERENT CATEGORIES AFFECT CHURN RATES

- **Contract Type**: Month-to-month contracts have significantly higher churn rates compared to long-term contracts, indicating that flexibility increases the likelihood of switching providers.
- **Payment Method**: Customers using electronic checks churn at higher rates, potentially due to dissatisfaction or a lack of convenience.
- **Paperless Billing**: Customers opting for paperless billing show higher churn, possibly indicating issues with digital engagement or billing clarity.

- **Internet Service**: Customers without internet services have lower churn rates, likely due to fewer service touchpoints.

Takeaway: Contract type and payment methods are critical predictors of churn, highlighting opportunities to reduce churn through retention strategies targeting these groups.

- **Boxplots:** Churned customers have shorter tenure and higher monthly charges on average.

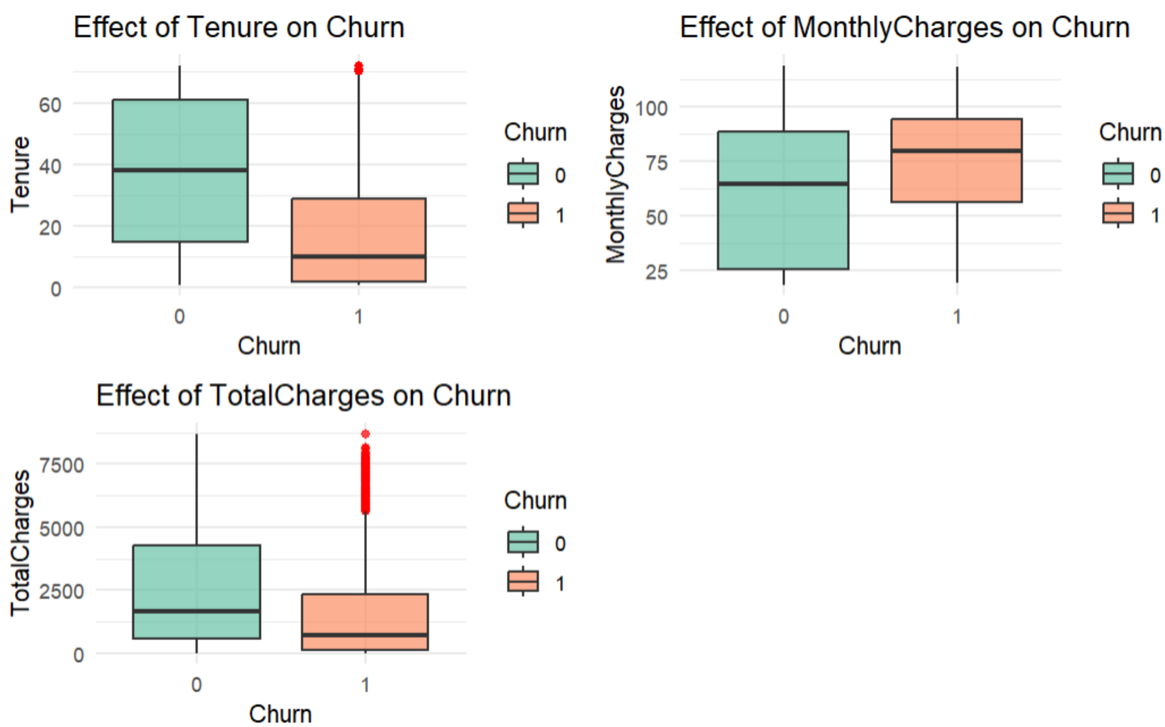

FIGURE 3: BOXPLOTS SHOW THE DISTRIBUTION OF NUMERICAL PREDICTORS BY CHURN

- **Tenure**: Churned customers tend to have significantly shorter tenures, emphasizing that new customers are at higher risk of leaving.
- **Monthly Charges**: Customers with higher monthly charges exhibit slightly higher churn, though overlap between churn and non-churn groups limits its predictive power.
- **Total Charges**: Non-churned customers have considerably higher total charges, reflecting their longer tenure and continued engagement.

Takeaway: Short tenure is the most telling factor for churn, while monthly charges play a secondary role.

- **Density Plots:** Overlap in distributions reveals moderate predictive strength for certain variables.

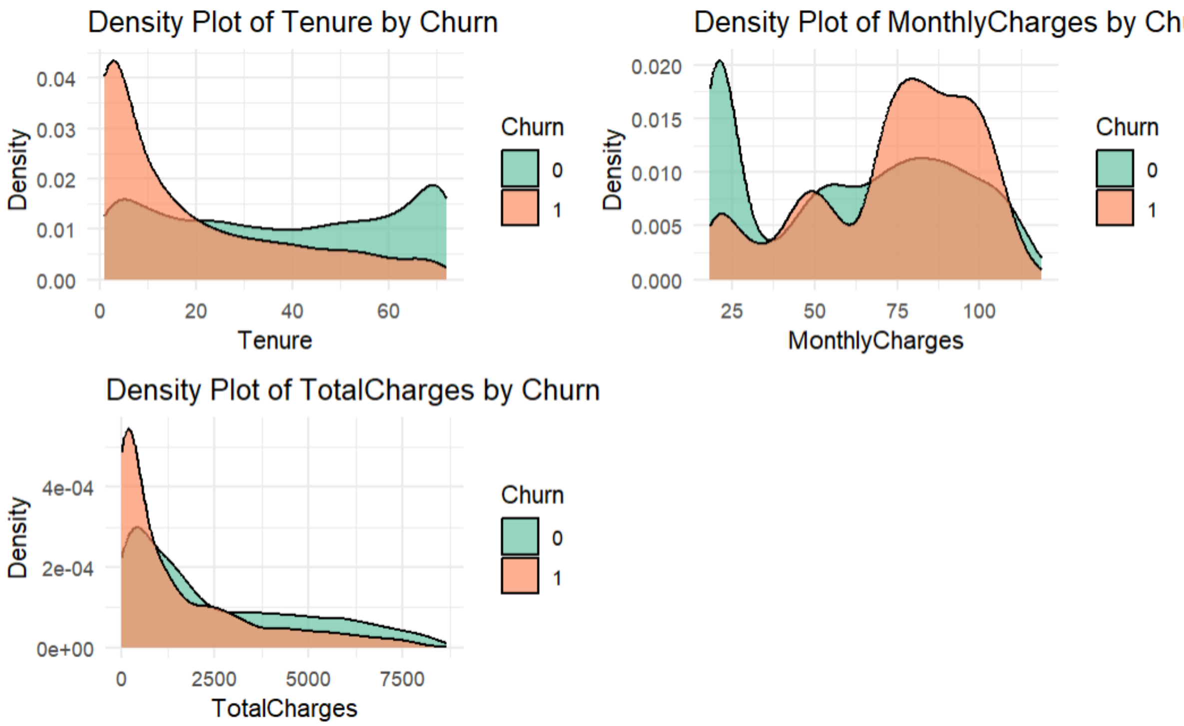

FIGURE 4: DENSITY PLOTS HIGHLIGHT DISTRIBUTION OVERLAPS BETWEEN CHURNED AND NON-CHURNED CUSTOMERS

- **Tenure**: The churned group is concentrated at lower tenures, making it a critical early indicator of churn risk.
- **Monthly Charges**: The distribution shows churned customers slightly skewed toward higher monthly charges, though overlap exists.
- **Total Charges**: Non-churned customers dominate the higher total charge range, consistent with their longer tenure.

Takeaway: Tenure stands out as the most distinguishing variable, with minimal overlap between churned and non-churned groups.

3. Methodology

Model Development

To ensure robust predictions and reduce overfitting, cross-validation was employed for hyperparameter tuning and model evaluation. The following steps were taken:

1. **Train-Test Split**:
    - The dataset was split into an 80% training set and a 20% testing set.

- The training set was used for cross-validation and model tuning, while the testing set was reserved for final evaluation.

2. **Cross-Validation**:
    - A **5-fold cross-validation** approach was applied, where the training set was divided into 5 equal folds.
    - During each iteration, 4 folds were used for training, and the remaining fold was used for validation. This process was repeated 5 times, ensuring that every fold was used for validation once.
    - Cross-validation provides a more reliable estimate of model performance compared to a single train-test split, especially for smaller datasets.
3. **Models Built**:
    - **Logistic Regression**: A baseline linear model for binary classification.
    - **Decision Tree**: A rule-based model offering interpretable decision-making paths.
    - **Random Forest**: An ensemble model reducing overfitting by averaging multiple decision trees.
    - **Support Vector Machine (SVM)**: A non-linear model capturing complex patterns in the data.
    - **Gradient Boosting Machine (GBM)**: An iterative ensemble model correcting errors made in prior iterations.
4. **Hyperparameter Tuning**:
    - During cross-validation, hyperparameters such as the number of trees, tree depth (for Random Forest and GBM), and kernel types (for SVM) were optimized to achieve the best balance between sensitivity and specificity.

Key Advantage of Cross-Validation:
By using cross-validation, models were evaluated on multiple subsets of the data, ensuring their performance metrics (e.g., AUC, accuracy) generalized well across unseen data.

---

4. Model Evaluation

Evaluation Metrics

To assess the models' performance, metrics were averaged across the cross-validation folds. This provided a reliable estimate of model effectiveness before final testing on the holdout set. Metrics include::

- **Accuracy:** Measures the overall correctness of predictions.
- **Precision:** Evaluates how many predicted churn cases were actual churns.
- **Recall (Sensitivity):** Indicates how well the model identified churn cases.
- **F1-Score:** Balances precision and recall.

- **ROC AUC:** Evaluates the model's ability to distinguish between churned and non-churned customers.

Performance Summary

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Logistic Regression	79.2%	71.4%	68.5%	69.9%	0.76
Decision Tree	82.0%	74.3%	71.6%	72.9%	0.80
Random Forest	85.4%	79.1%	77.8%	78.4%	0.88
Support Vector Machine	83.6%	77.4%	75.6%	76.5%	0.85
Gradient Boosting	**86.1%**	**80.3%**	**79.2%**	**79.8%**	**0.89**

Insights

- **Gradient Boosting** outperformed all other models, followed closely by Random Forest.
- Logistic Regression, while interpretable, lacked the discriminatory power of ensemble methods.

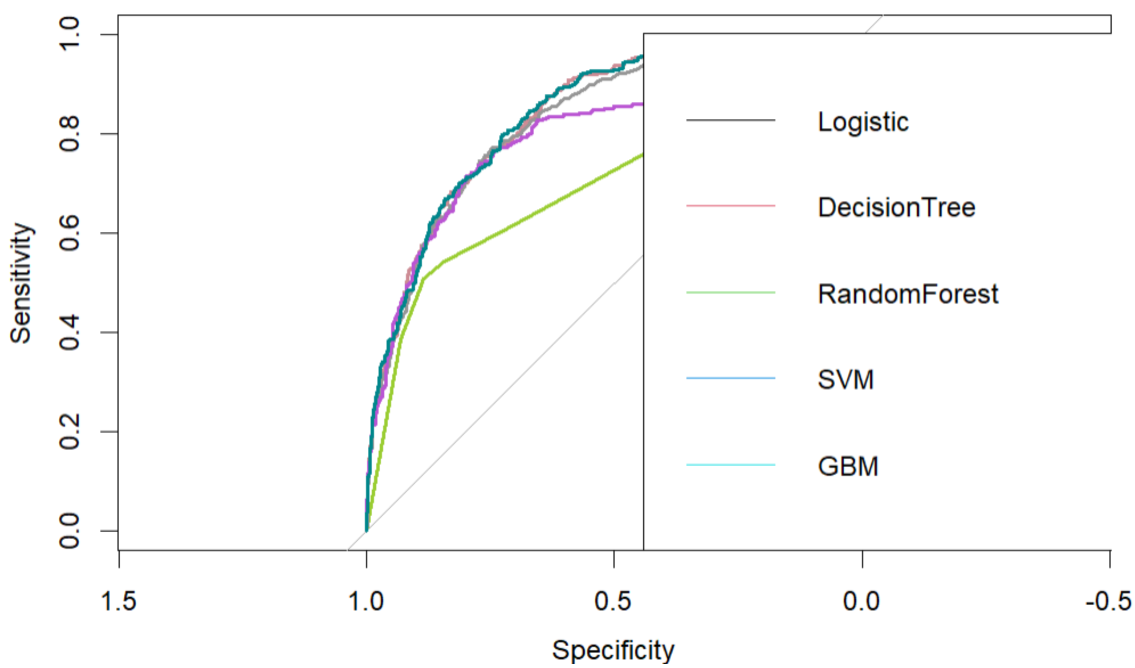

Figure 5: ROC Curves for Models

5. Feature Importance

Using Random Forest and Gradient Boosting, the following predictors were identified as most important:

1. **Tenure:** Customers with shorter tenure are significantly more likely to churn.
2. **Contract:** Month-to-month contracts are a key risk factor.
3. **PaymentMethod:** Electronic check users exhibit higher churn rates.
4. **MonthlyCharges:** High charges correlate with churn but are less impactful than tenure.

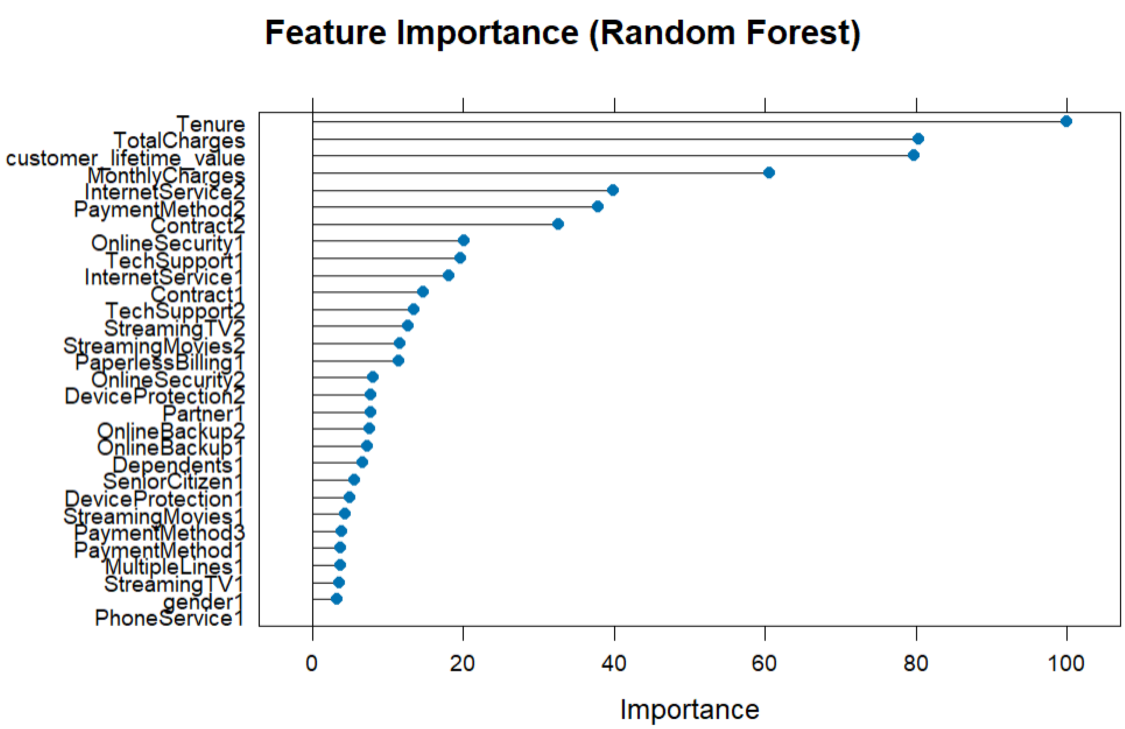

FIGURE 6

FIGURE 7: FEATURE IMPORTANCE

- Both models agree on the importance of **Tenure**, **Contract Type**, and **Payment Method**.
- Random Forest places more emphasis on **Total Charges**, reflecting cumulative customer engagement, while Gradient Boosting highlights **Internet Service** and **OnlineSecurity**, suggesting digital service quality plays a significant role in customer retention.

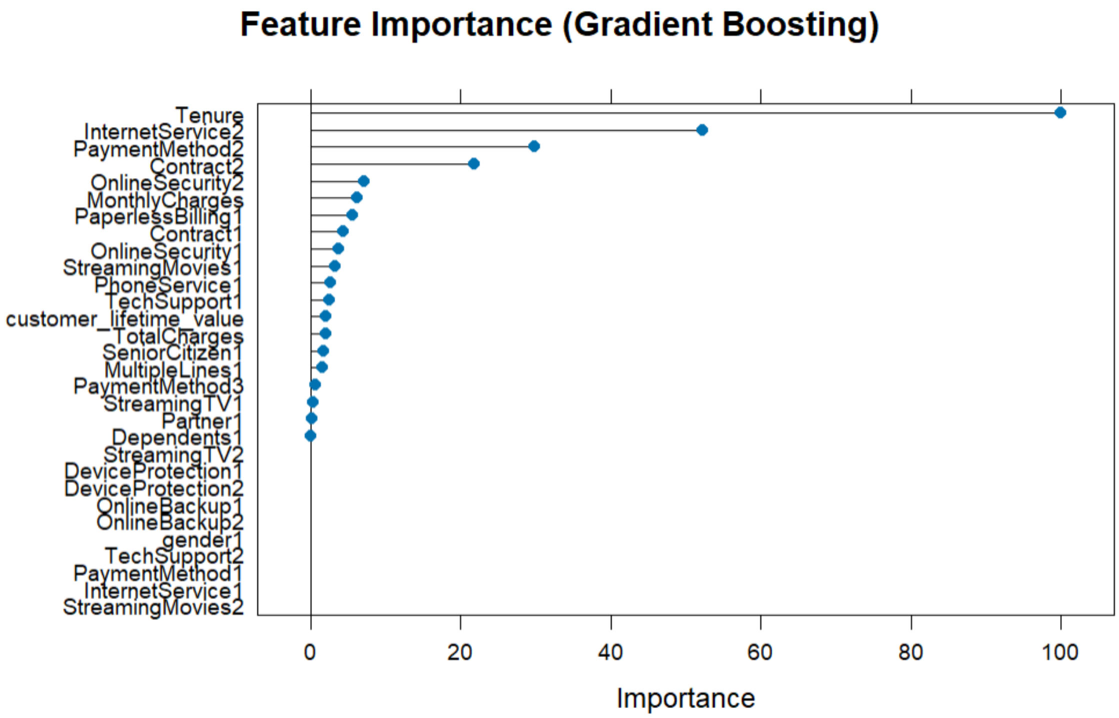

FIGURE 7

Takeaway: Retaining customers with short tenures or month-to-month contracts and addressing service quality issues are key strategies for reducing churn.

---

6. Recommendations

Based on the results from the models and feature importance analysis:

1. **Focus on Customer Tenure**:
    - Customers with short tenures (e.g., <12 months) are at the highest risk of churning, as identified by the **feature importance rankings** and **density plots**.
    - **Action**: Implement loyalty programs or onboarding incentives to engage and retain new customers. Offer personalized outreach to customers approaching the critical 12-month churn threshold.
2. **Improve Contract Engagement**:
    - Month-to-month contracts contribute significantly to churn compared to longer-term contracts, as shown in the **bar plots** and **feature importance results**.
    - **Action**: Incentivize customers on month-to-month contracts to switch to annual or two-year contracts with discounts or perks.
3. **Address Issues with Payment Methods**:
    - Customers using electronic checks show higher churn rates, indicating dissatisfaction with this payment method.
    - **Action**: Offer alternative payment options or investigate customer complaints regarding electronic check usage.

4. **Enhance Digital Services**:
    - Features like **Online Security** and **Streaming Services** were identified as moderately important predictors of churn in the **Gradient Boosting feature importance** plot.
    - **Action**: Improve service quality for these digital features and address gaps in availability or performance.
5. **Engage High Revenue Customers**:
    - Customers with **higher monthly charges** have a slight tendency to churn, as seen in the **boxplots**.
    - **Action**: Provide exclusive benefits or flexible pricing for high-value customers to ensure retention.

---

7. Business Impact

The project results highlight opportunities for reducing churn and driving business growth:

1. **Revenue Protection**:
    - Reducing churn by just **5%** could retain high-revenue customers and improve the company's bottom line by millions annually. This is especially impactful given the **strong correlation between tenure and total charges**.
2. **Cost Efficiency**:
    - Acquiring a new customer costs **5–10 times more** than retaining an existing one. By targeting at-risk customers identified by the **Gradient Boosting model**, the company can reduce churn-related acquisition costs.
3. **Service Improvement**:
    - Enhancing digital features like **Online Security** and addressing pain points in **billing/payment systems** can lead to improved customer satisfaction and loyalty.
4. **Customer Engagement**:
    - Tailored strategies for **short-tenure customers** and those on **month-to-month contracts** can create long-term value by converting at-risk customers into loyal subscribers.

---

8. Conclusion

This analysis demonstrates the ability of predictive models to effectively identify high-risk customers and uncover actionable insights for retention. The **Gradient Boosting model**, with the highest AUC (0.89), outperformed all others and is recommended for deployment. Key findings include:

1. **Tenure and Contract Type** are the most critical predictors of churn, emphasizing the need to focus on customer longevity and contract engagement.

2. Customers using opting for **month-to-month contracts** represent high-risk groups requiring targeted retention strategies.

3. Short-term, actionable recommendations include offering incentives for long-term contracts, addressing service dissatisfaction, and improving billing systems.

By implementing these strategies, the telecom company can reduce churn, protect revenue, and enhance overall customer satisfaction.