Business Analytics with R
Group Project

Customer Churn Analysis in the Telecom Industry

Avlani, Vedansh H

Tran, Bao Gia

Thandayutham, Nagarajan

Wen, Xue

Summary

Customer churn is a critical challenge in the telecom industry, directly impacting revenue and customer acquisition costs. This report uses a robust dataset and machine learning models to predict churn and identify factors influencing it. The analysis highlights actionable insights and strategies to improve customer retention, reduce churn, and enhance long-term growth.

---

1. Objective

The primary objective of this analysis is to predict customer churn and identify the factors contributing to it. Insights gained aim to inform proactive measures for enhancing customer retention and mitigating revenue loss.

---

2. Data Overview

2.1 Dataset Description

The dataset contains **7,043 customer records** and 21 features, divided into:

- **Demographics**: Gender, Senior Citizen, Partner, Dependents.
- **Service Details**: Tenure, Phone Service, Internet Service, Online Security, etc.
- **Payment Details**: Contract Type, Monthly Charges, Total Charges.
- **Churn Indicator**: Whether a customer churned (Yes/No).

2.2 Key Characteristics

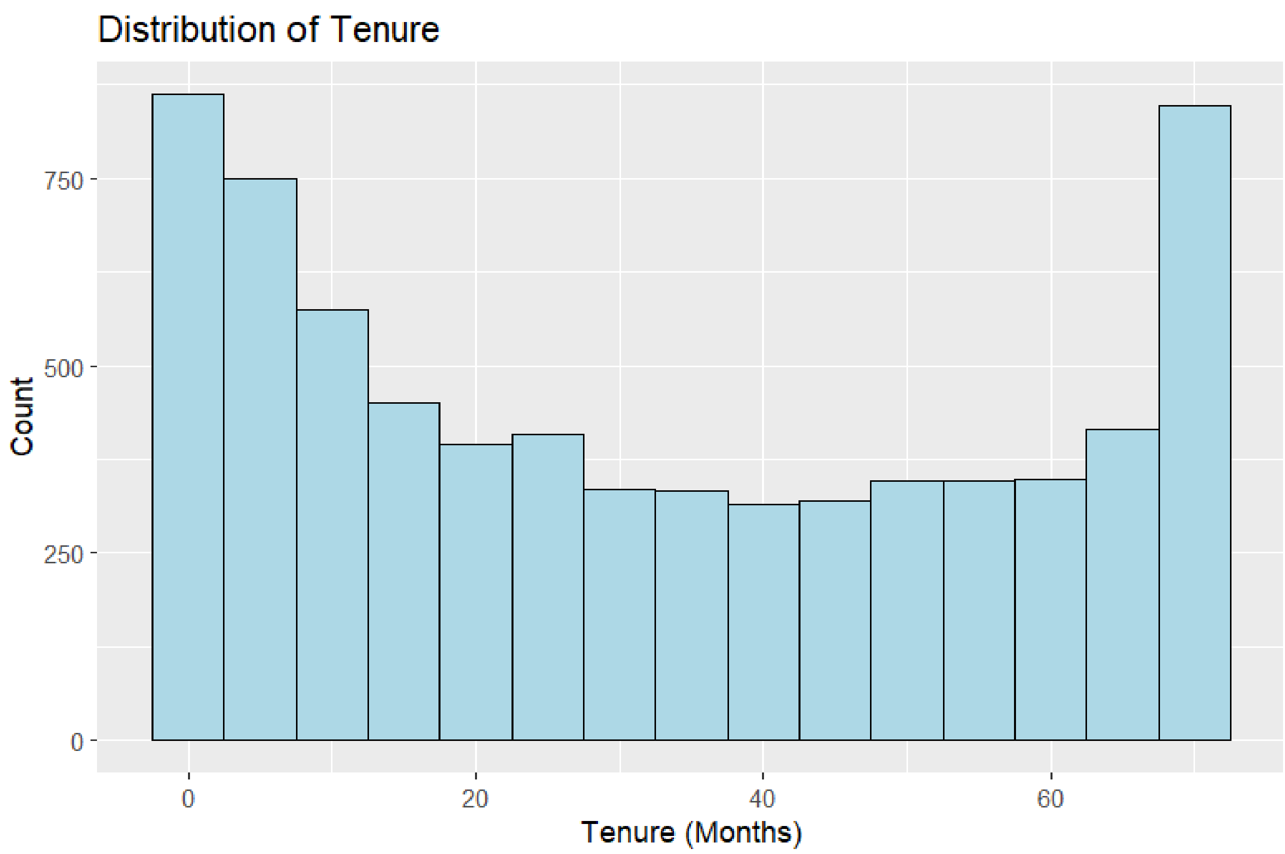
1. **Gender Distribution**: Balanced representation between male and female customers.
2. **Contract Types**: High churn observed in month-to-month contracts.
3. **Tenure**: Bimodal distribution showing early churn within the first few months.

---

3. Exploratory Data Analysis (EDA)

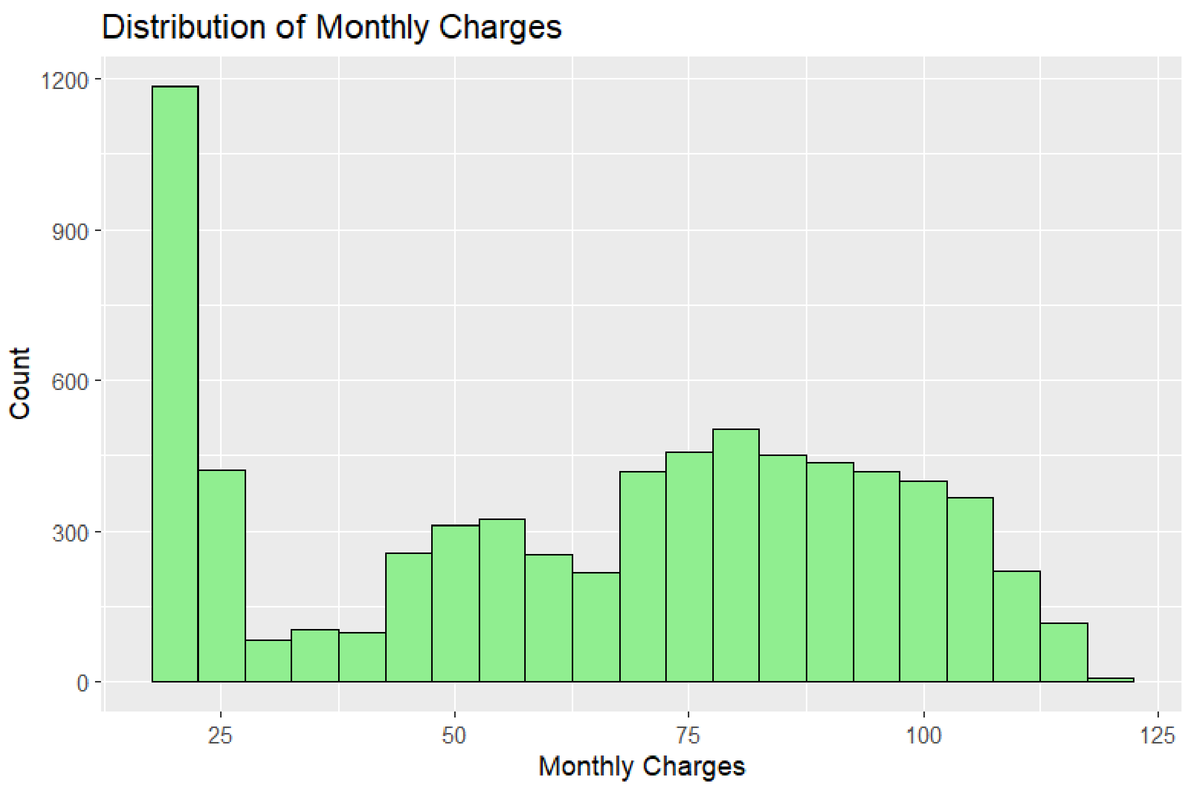
3.1 Tenure Distribution

- **Insights**: A bimodal distribution shows:
  - Early churn within the first 3–6 months.
  - Long-term retention after surpassing this critical period.
- **Recommendation**: Improve onboarding processes and focus on initial engagement to reduce early churn.



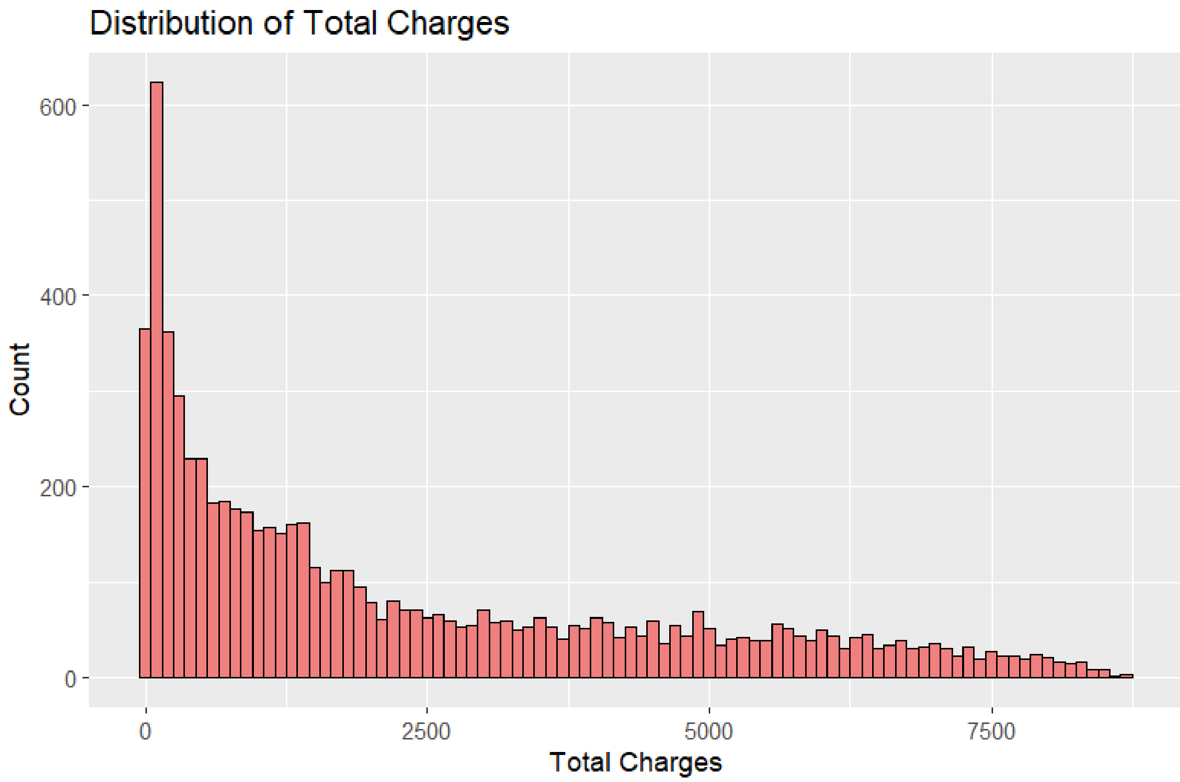
3.2 Monthly Charges Distribution

- **Insights**: Right-skewed; most customers opt for lower-tier plans, with higher churn rates among high-paying customers.
- **Recommendation**: Offer loyalty discounts or bundle premium services to enhance value perception.



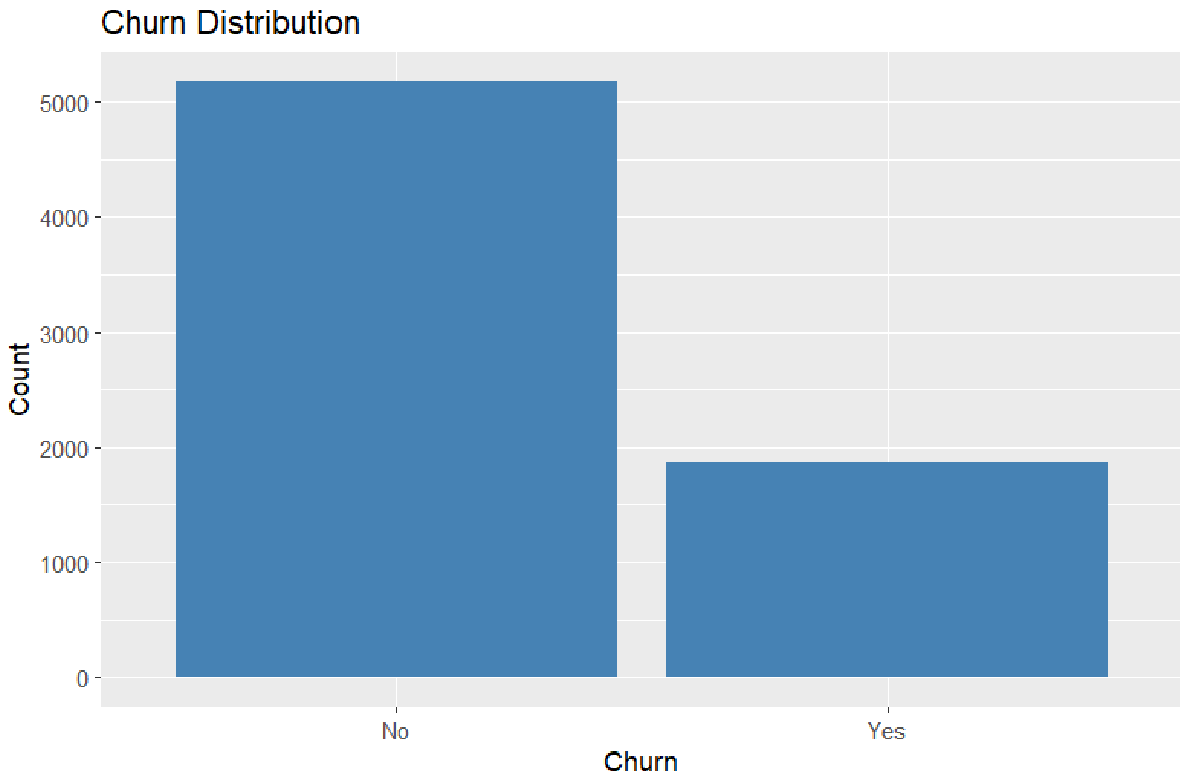
3.3 Total Charges Distribution

- **Insights**: Highly skewed; early churners contribute to lower total charges.
- **Recommendation**: Implement targeted retention strategies for new customers to extend tenure and maximize lifetime value.



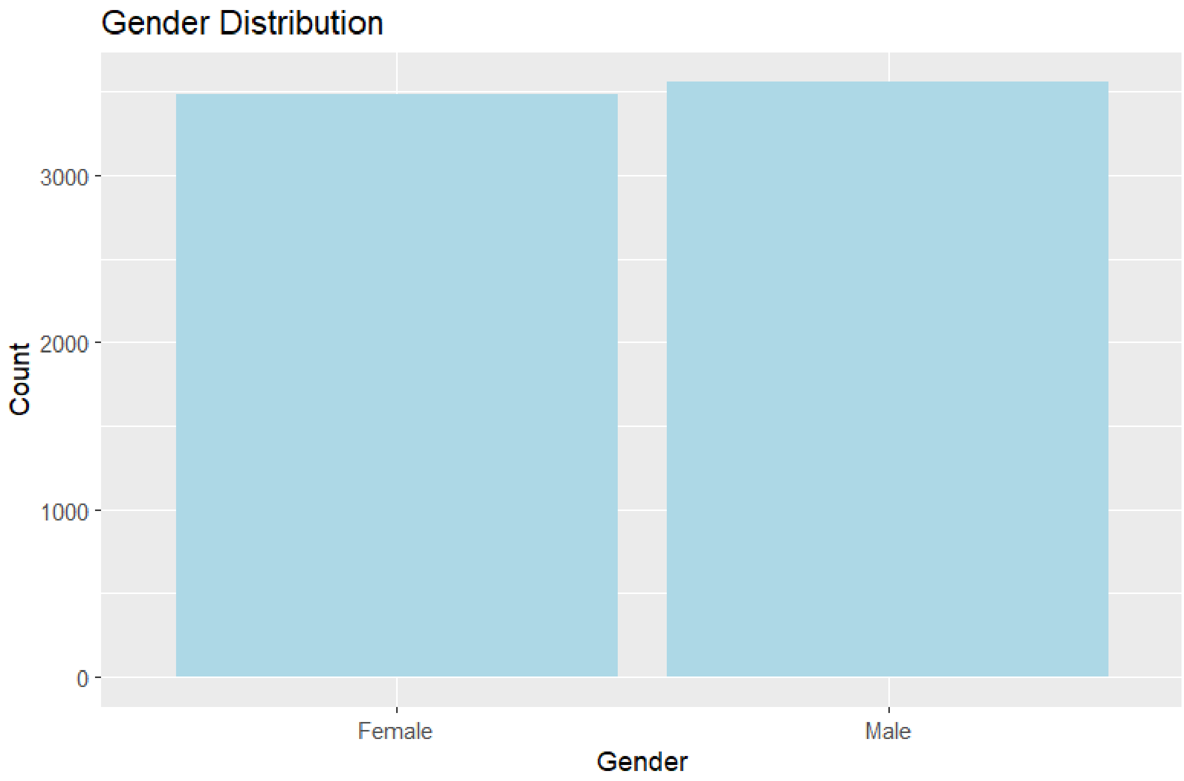
3.4 Churn Distribution

- **Insights**: Month-to-month contracts and higher charges correlate strongly with churn.
- **Recommendation**: Identify and address dissatisfaction in these groups.



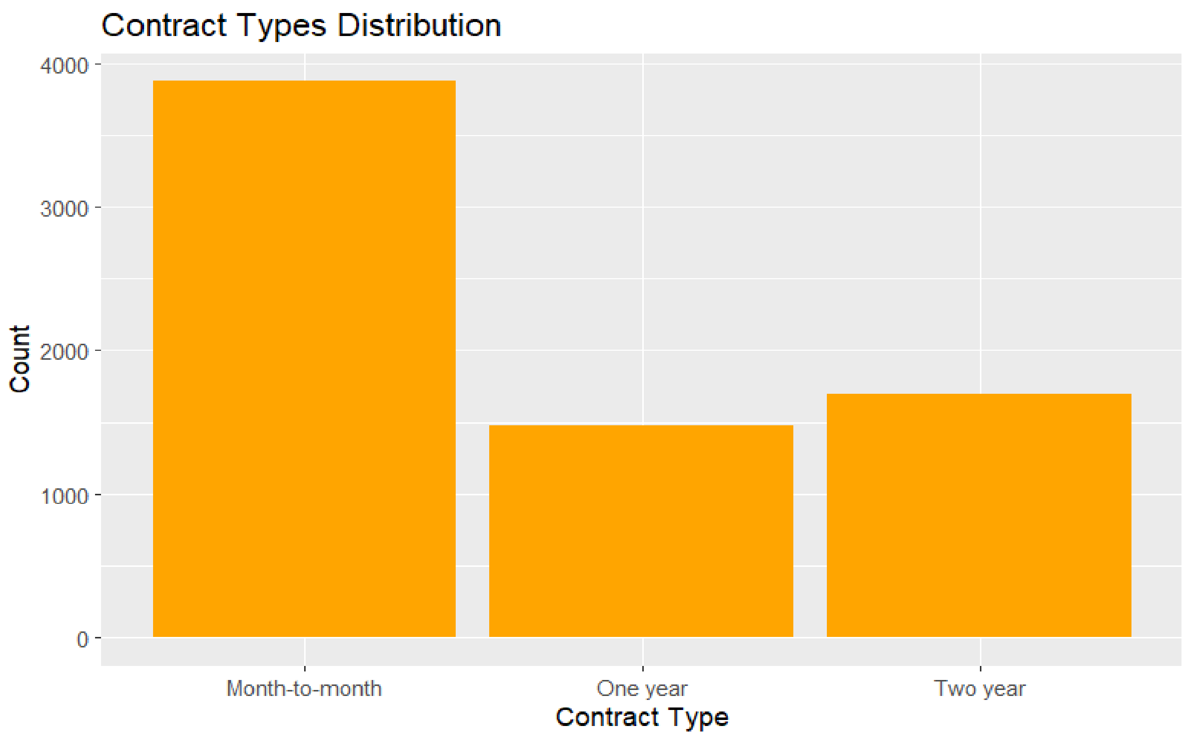
3.5 Gender Distribution

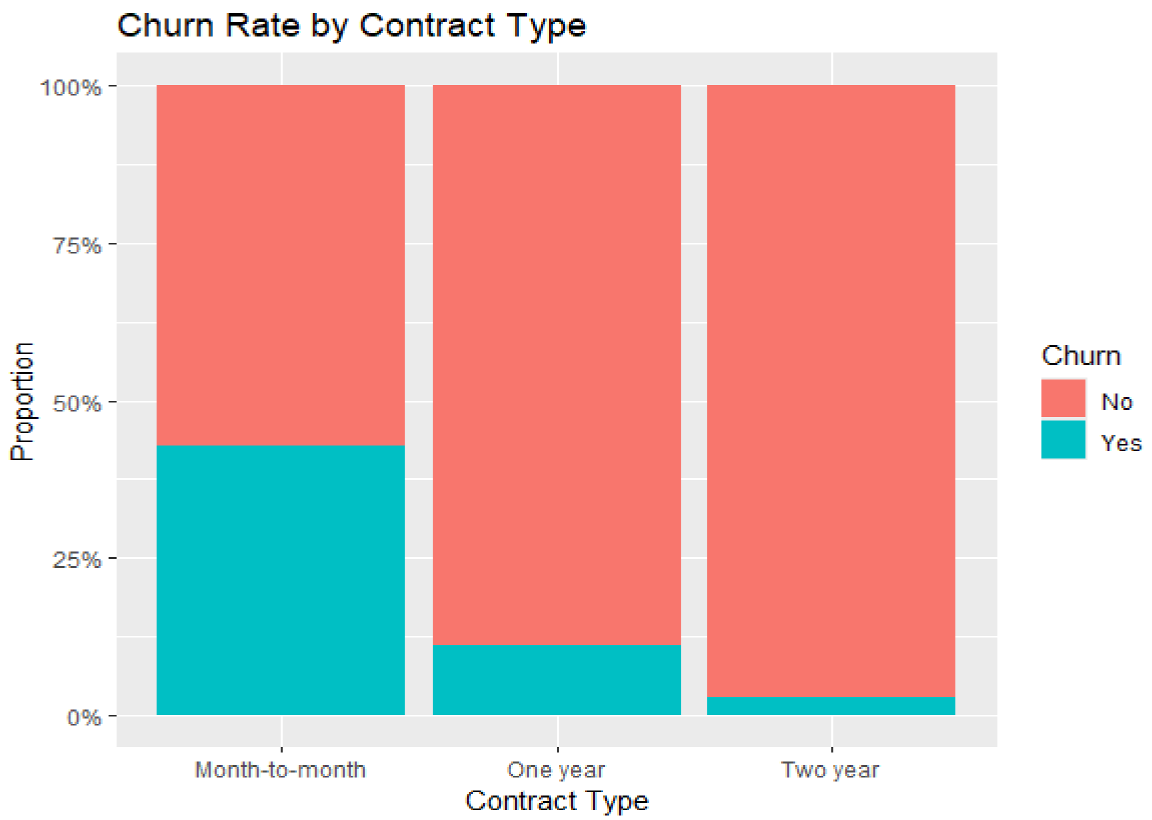
- **Insights**: Gender balance indicates broad service appeal, but differences in churn patterns by gender may exist.
- **Recommendation**: Explore gender-specific churn drivers for personalized interventions.



3.6 Contract Types

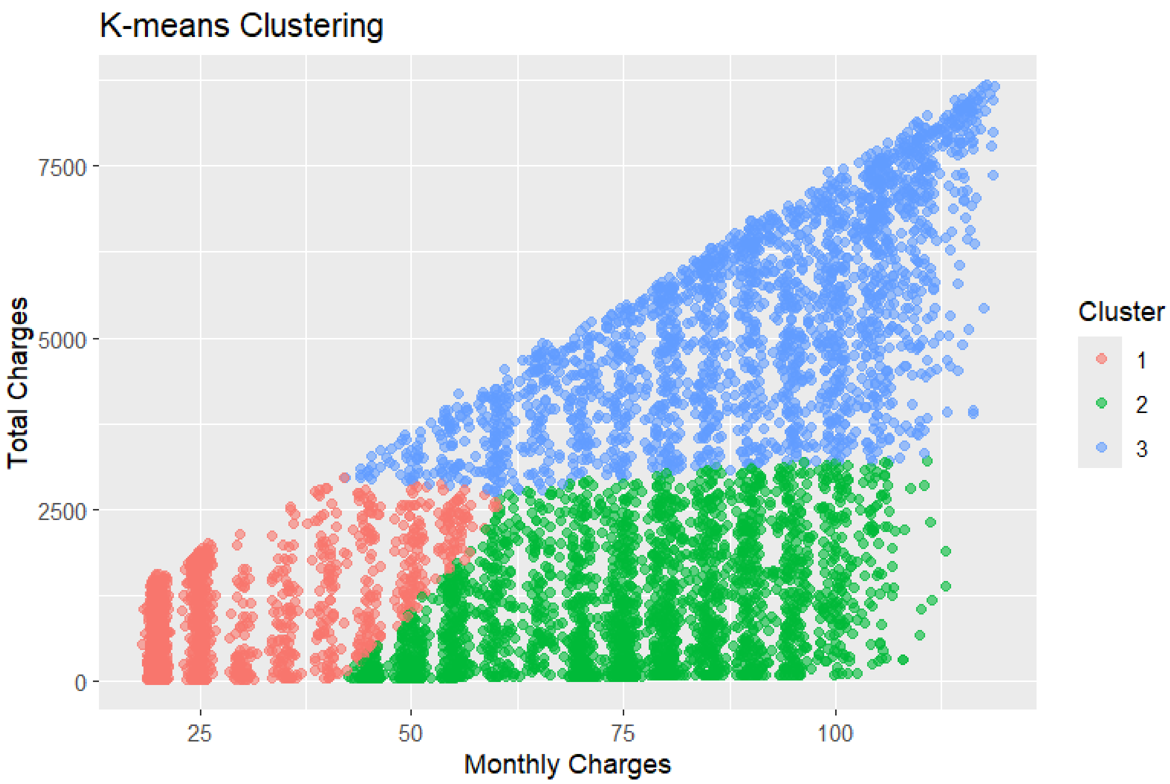
- **Insights**: Month-to-month contracts exhibit significantly higher churn rates.
- **Recommendation**: Offer discounts or benefits for switching to longer-term contracts.





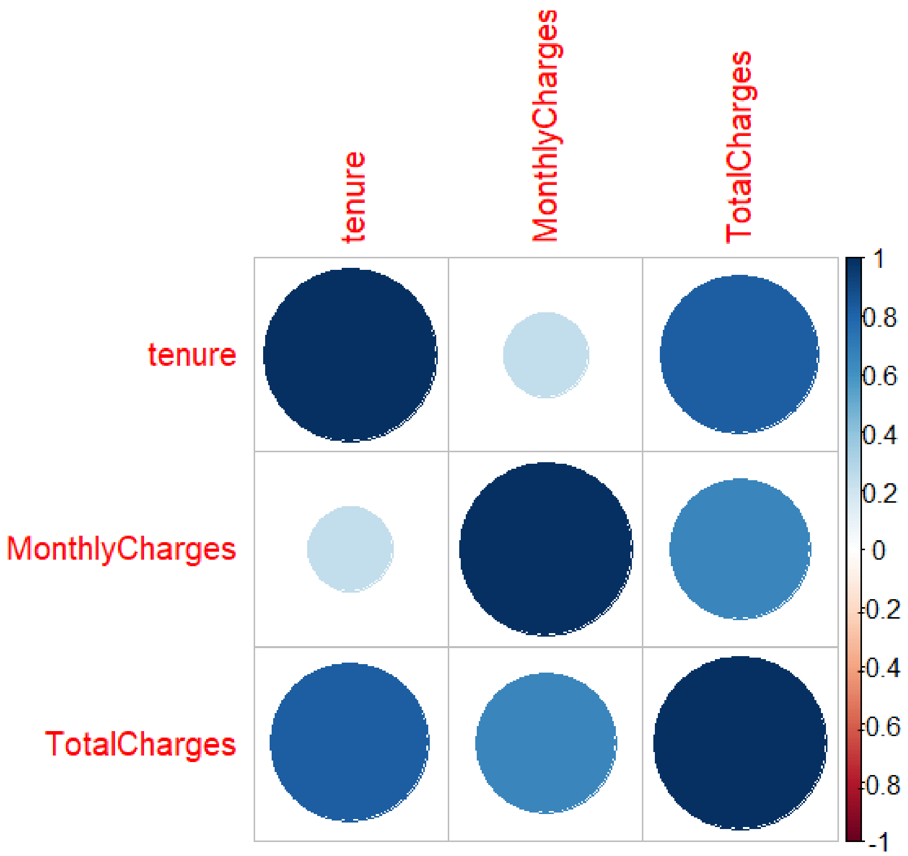
3.7 Clustering Analysis

- **Insights**:
  - Cluster 1: Low charges and tenure (high churn risk).
  - Cluster 2: Moderate charges and tenure.
  - Cluster 3: High charges and tenure (low churn risk).
- **Recommendation**: Focus retention efforts on Cluster 1, addressing dissatisfaction and upselling opportunities.



3.7 Correlation Analysis

- **Insights:**
  - Strong positive correlation between tenure and total charges: As tenure increases, total charges grow due to cumulative monthly payments.
  - Positive correlation between monthly charges and total charges: High monthly charges lead to higher total charges over time.
  - Weak correlation between tenure and monthly charges: Monthly charges alone do not determine customer longevity.
- **Implication:** Long-term retention depends on factors like customer satisfaction, perceived value, and service quality rather than pricing alone.
- **Recommendation:** Enhance customer satisfaction and perceived value to retain customers beyond initial periods.



4. Methodology

4.1 Data Preprocessing

- Missing values in TotalCharges handled.
- Non-predictive features removed (e.g., Customer ID).
- Variables converted to appropriate formats.

4.2 Class Balancing

- The ROSE technique was applied to balance churn distribution, ensuring model fairness.

4.3 Models Evaluated

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Neural Network

---

5. Model Evaluation

5.1 Logistic Regression

- **Accuracy**: **75.69%**
- **Confusion Matrix:**

```
log_pred_class  No Yes
           No  771  70
           Yes 272 294
```

- **Strengths**: High interpretability; significant predictors include tenure, contract type, and monthly charges.
- **Use Case**: Effective for baseline churn risk assessment.

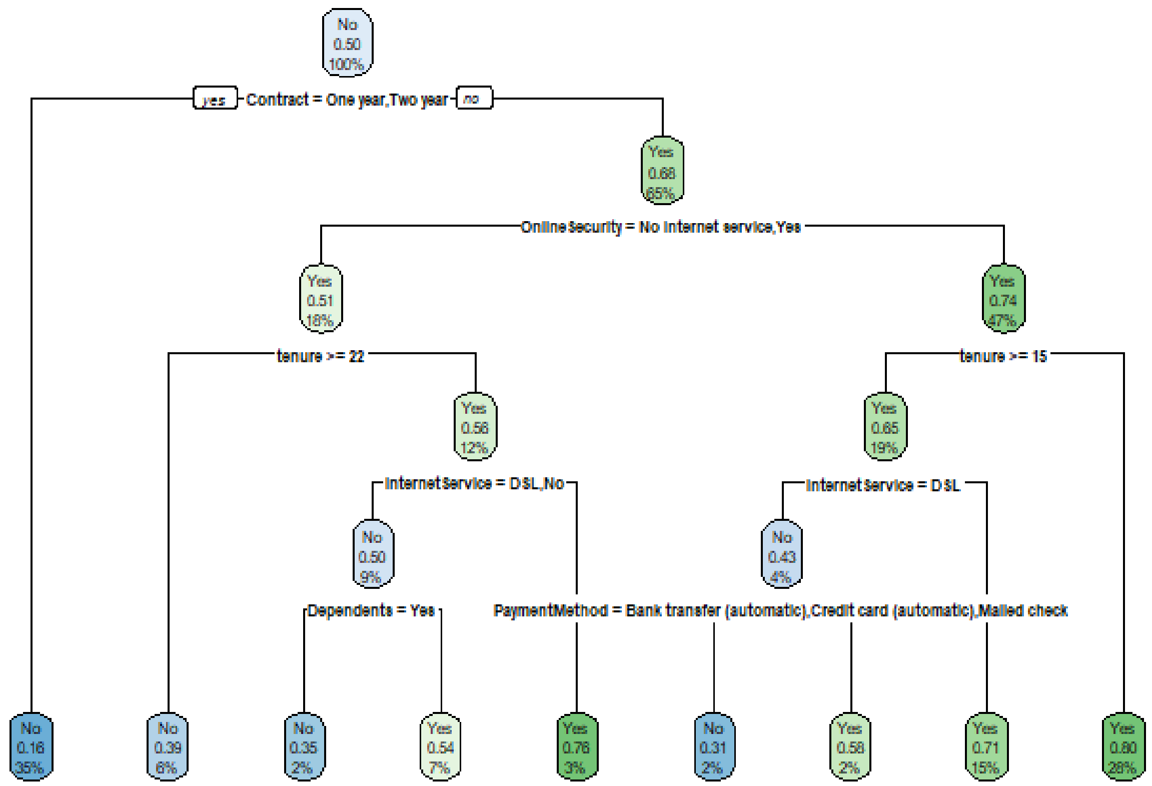
5.2 Decision Tree

- **Accuracy**: **72.64%**
- **Confusion Matrix:**

```
          Reference
Prediction  No Yes
       No  729  71
       Yes 314 293
```

- **Strengths**: Clear decision rules; tenure and contract type are key drivers.
- **Limitations**: Slightly lower accuracy; prone to overfitting without pruning.
- **Use Case**: Valuable for visualizing customer behaviour.

Decision Tree for Churn Prediction

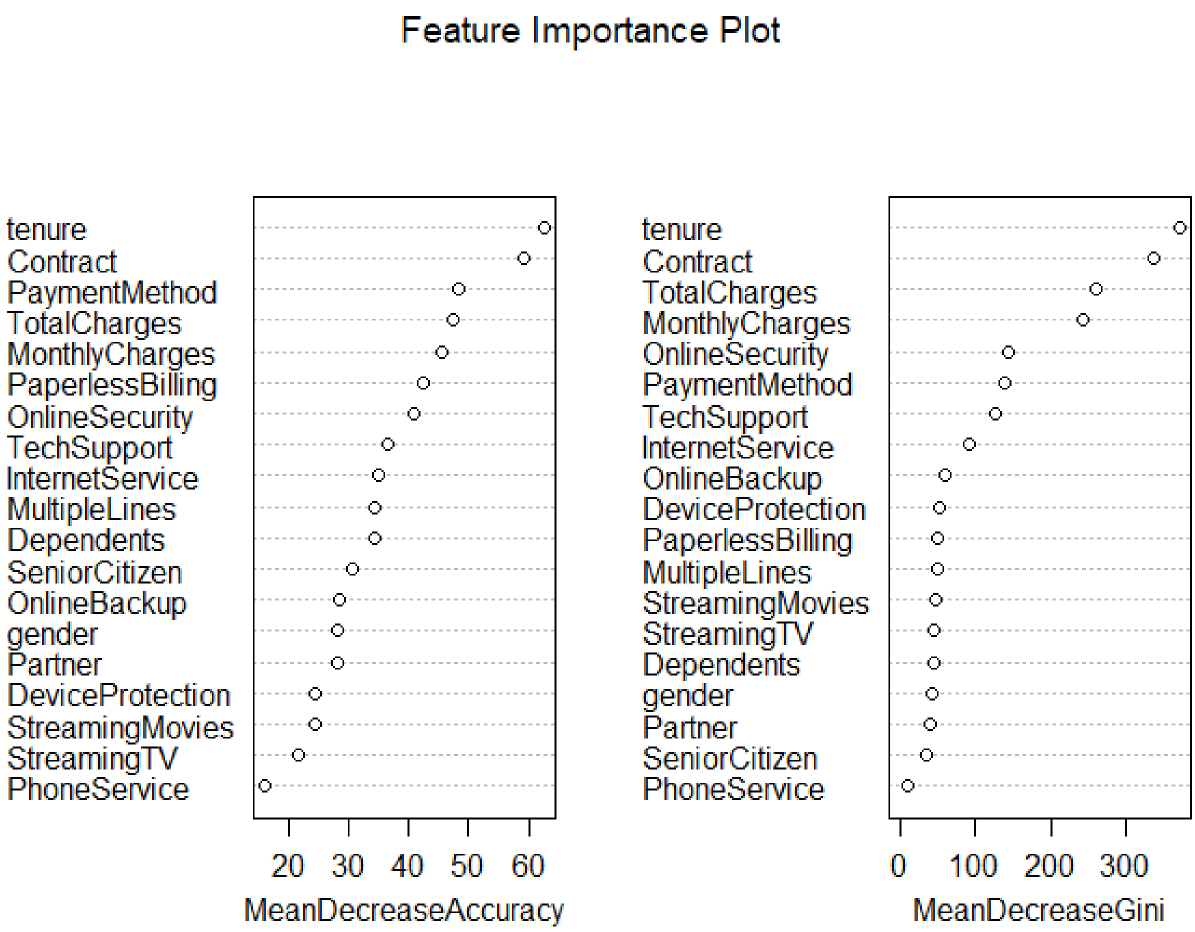


5.3 Random Forest

- **Accuracy**: **73.56%**
- **Confusion Matrix:**

```
          Reference
Prediction  No Yes
       No  736  65
       Yes 307 299
```

- **Strengths**: Robust and handles non-linear relationships effectively.
- **Limitations**: Computationally intensive; lower interpretability.
- **Use Case**: Best for feature importance analysis and handling complex data.

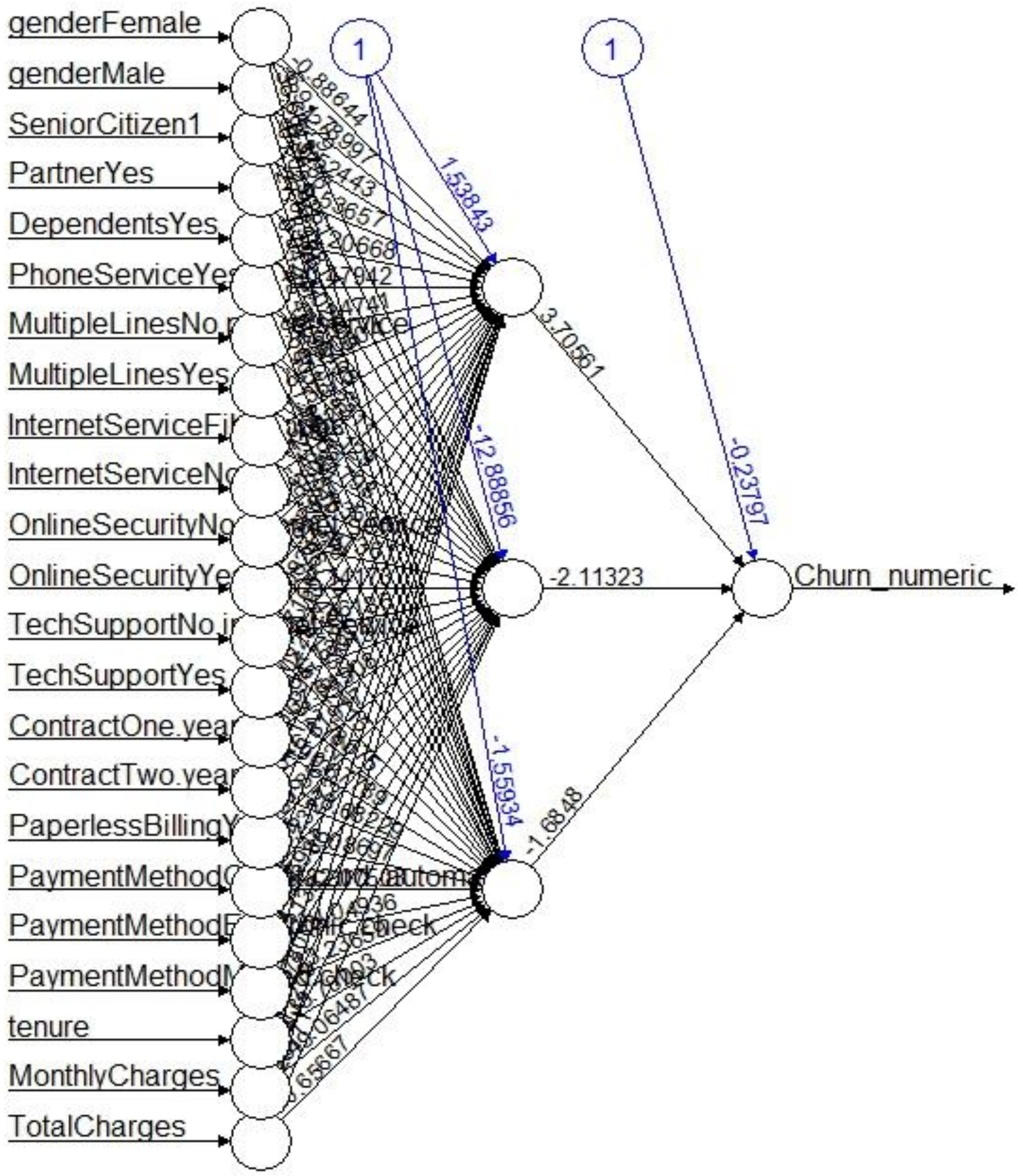


5.4 Neural Network

- **Accuracy**: **72.14%**
- **Confusion Matrix:**

```
nn_pred_class   No Yes
           No  722  71
           Yes 321 293
```

- **Strengths**: Captures non-linear patterns in data.
- **Limitations**: Requires feature scaling and encoding; lower interpretability.
- **Use Case**: Advanced predictive analysis for larger datasets.



6. Model Comparison

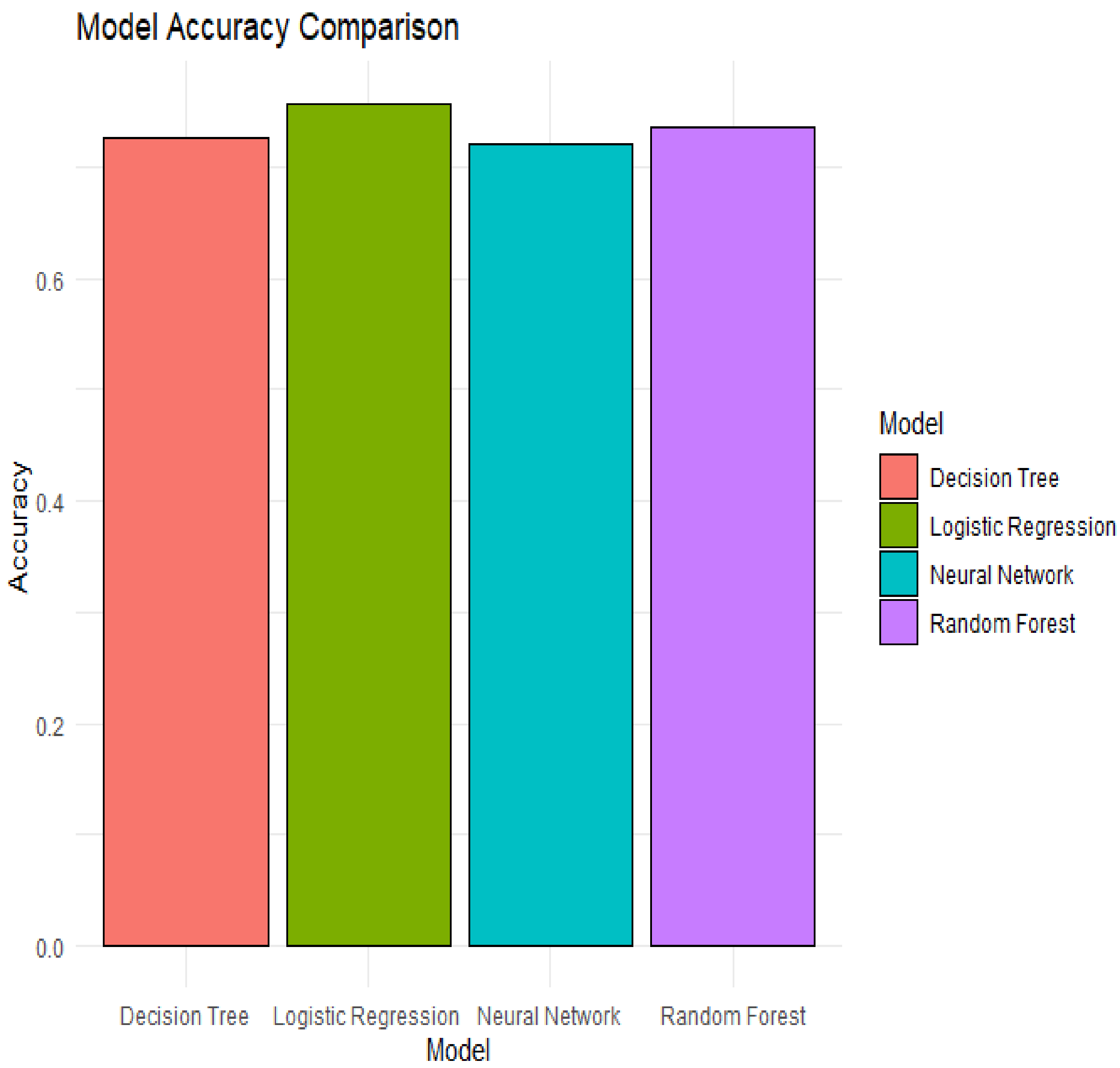
- Logistic Regression performed best overall with 75.69% accuracy
- Random Forest showed highest specificity (82.14%)
- All models showed relatively consistent performance around 72-76% accuracy

Model Finalized: Random Forest

Reasons for finalizing Random Forest:

1. Stability: This model shows more stable predictions, which are important in business contexts where consistency and reliability matter.
2. Lower Error: The graph appears to exhibit fewer deviations from the expected values, indicating a lower error rate compared to Model 1, which has more significant fluctuations.
3. Generalization: A model with fewer spikes and more steady behavior is often better at generalizing to unseen data, as it isn't overfitting to noise or outliers.

Model	Accuracy	Key Strength
Logistic Regression	75.69%	Interpretability
Decision Tree	72.64%	Rule-based insights
Random Forest	73.56%	Feature importance
Neural Network	72.14%	Non-linear modeling


Model Accuracy Comparison
Accuracy
0.0
0.2
0.4
0.6
Decision Tree
Logistic Regression
Neural Network
Random Forest
Model
Model
Decision Tree
Logistic Regression
Neural Network
Random Forest

```
Model                  Accuracy    Sensitivity    Specificity
--------------------------------------------------------------
Logistic Regression   0.7569      0.7392         0.8077
Random Forest         0.7356      0.7057         0.8214
Decision Tree         0.7264      0.6989         0.8049
```

```
Model                 True Neg    False Pos    False Neg    True Pos
-----------------------------------------------------------------
Logistic Regression     771         272          70          294
Random Forest           736         307          65          299
Decision Tree           729         314          71          293
```

```
Model                 Accuracy    Architecture    Comments
----------------------------------------------------------------
Logistic Regression   0.7569      N/A             Best overall
Random Forest         0.7356      500 trees       Good balance
Decision Tree         0.7264      Max depth 8     Interpretable
Neural Network        ~0.73*      3 hidden units  Complex model
```

7. Recommendations

- Contract & Pricing

  Target month-to-month contract customers for conversion to long-term contracts, as they represent 43% of all churners and have a significantly higher churn rate

- Service Bundling

  Bundle Tech Support with Fiber Optic Service packages, as data shows lack of Tech Support increases churn by 35% and Fiber Optic users without support have 42% higher churn rates

- High-Risk Identification

  Prioritize intervention for customers with less than 12 months tenure on month-to-month contracts, as this segment shows the highest correlation with churn in our analysis

- Revenue Protection

  Focus retention efforts on the identified 294 high-risk customers representing $26,460 in monthly revenue, as preventing their churn could save $317,520 annually

- Model Application

  Implement the Random Forest model (82.14% Specificity) as the primary prediction tool, focusing on the identified loyal customers for upselling while monitoring the high-risk cases

8. Conclusion

The customer churn analysis has highlighted key factors influencing churn, including contract type, monthly charges, and customer tenure. Month-to-month contracts show significantly higher churn rates, suggesting that contract flexibility may contribute to customer turnover. To mitigate this, the company could introduce incentives for customers to commit to longer-term contracts, such as discounts or added benefits, which would reduce churn by encouraging greater customer loyalty. Additionally, customers with higher monthly charges appear more prone to churn, potentially due to cost dissatisfaction. Implementing strategies to enhance perceived value for premium customers, such as exclusive services or personalized support, could improve retention among high-paying customers.

The distribution of tenure indicates that churn is most likely to occur in the early stages of the customer lifecycle, highlighting the importance of initial engagement. Strategies like personalized onboarding, proactive customer support during the first few months, and loyalty rewards for continued subscription could strengthen early customer relationships, leading to longer tenures and increased lifetime value. The clustering analysis further suggests that targeted retention efforts could focus on specific customer segments identified as higher-risk, particularly those with lower monthly charges, who may benefit from tailored upsell offers.

Overall, this analysis provides a strong foundation for a data-driven retention strategy. By leveraging insights into contract type, pricing, and customer segmentation, the company can take proactive steps to reduce churn. Implementing predictive models to monitor churn risk and using targeted retention interventions can improve customer satisfaction and foster long-term growth.
