Strategic Analysis of Customer Churn: A Machine Learning Approach

A PROJECT REPORT SUBMITTED BY:

Y. C. K. Arachchi

BSc in Physical Science (Statistics & Computer Science)

Faculty of Science

University of Peradeniya

Sri Lanka

01/02/2026

Abstract

Customer Churn Prediction using Machine Learning: A Random Forest Approach

Customer attrition, or "churn," represents a significant revenue challenge for the telecommunications industry, where the cost of acquiring new customers far exceeds that of retaining existing ones. This project investigates the factors contributing to customer churn and develops a predictive model to identify high-risk subscribers. Using the "Telco Customer Churn" dataset comprising 7,043 observations, the study employed the R programming language to perform data preprocessing, feature engineering, and exploratory data analysis (EDA). A **Random Forest** classification algorithm was trained and optimized using 5-fold cross-validation. The final model achieved an accuracy of **78.6%** on an unseen test set, successfully distinguishing between loyal and churning customers. Feature importance analysis revealed that **Contract Type**, **Tenure**, and **Monthly Charges** are the most significant predictors of attrition. The findings suggest that customers on month-to-month contracts and those in their first year of service are the most vulnerable segments. Consequently, this report recommends strategic interventions, including incentivized long-term contracts and enhanced onboarding programs, to mitigate churn and improve customer lifetime value.

Keywords: *Machine Learning, Customer Churn, Random Forest, R Programming, Business Analytics, Predictive Modeling.*

1. Executive Summary

Objective In the competitive telecommunications sector, customer acquisition costs are estimated to be 5–25 times higher than retention costs. The objective of this project was to leverage machine learning to predict customer attrition (churn) and identify key risk factors driving customers away.

Methodology We analyzed a dataset of 7,043 customer records using the R programming language. The analysis followed a standard data science pipeline:

1. **Data Preprocessing:** Cleaning missing values and encoding categorical variables.
2. **Exploratory Data Analysis (EDA):** Visualizing relationships between churn and account features.
3. **Machine Learning:** Training a **Random Forest Classifier** on 75% of the data.
4. **Evaluation:** Testing the model on a 25% unseen hold-out set.

Key Findings

- The predictive model achieved an **Accuracy of 78.6%**.
- The primary drivers of churn are **Contract Type**, **Tenure**, and **Monthly Charges**.
- Customers on **Month-to-Month contracts** are the highest risk group.
- Churn likelihood is highest within the first **12 months** of service.

Strategic Recommendation To reduce churn, the business should focus on migrating month-to-month customers to 1-year contracts through targeted incentives and implement a robust "First Year" onboarding program to support new users.

2. Data & Methodology

2.1 Data Overview The analysis utilized the IBM "Telco Customer Churn" dataset, comprising 7,043 observations and 21 variables. Key features included:

- **Demographics:** Gender, Senior Citizen status, Partner/Dependents.
- **Services:** Phone, Internet (DSL/Fiber), Tech Support, Streaming.
- **Account Info:** Contract Type, Payment Method, Monthly Charges.

2.2 Data Cleaning Raw data often contains inconsistencies. We performed the following preprocessing steps in R:

- Removed 11 rows containing missing `TotalCharges` values to ensure data integrity.
- Removed the `customerID` column as it offered no predictive value.
- Converted categorical variables (e.g., "Yes/No") into Factors for machine learning compatibility.

2.3 Exploratory Data Analysis (EDA) Visual analysis revealed critical insights into customer behavior.

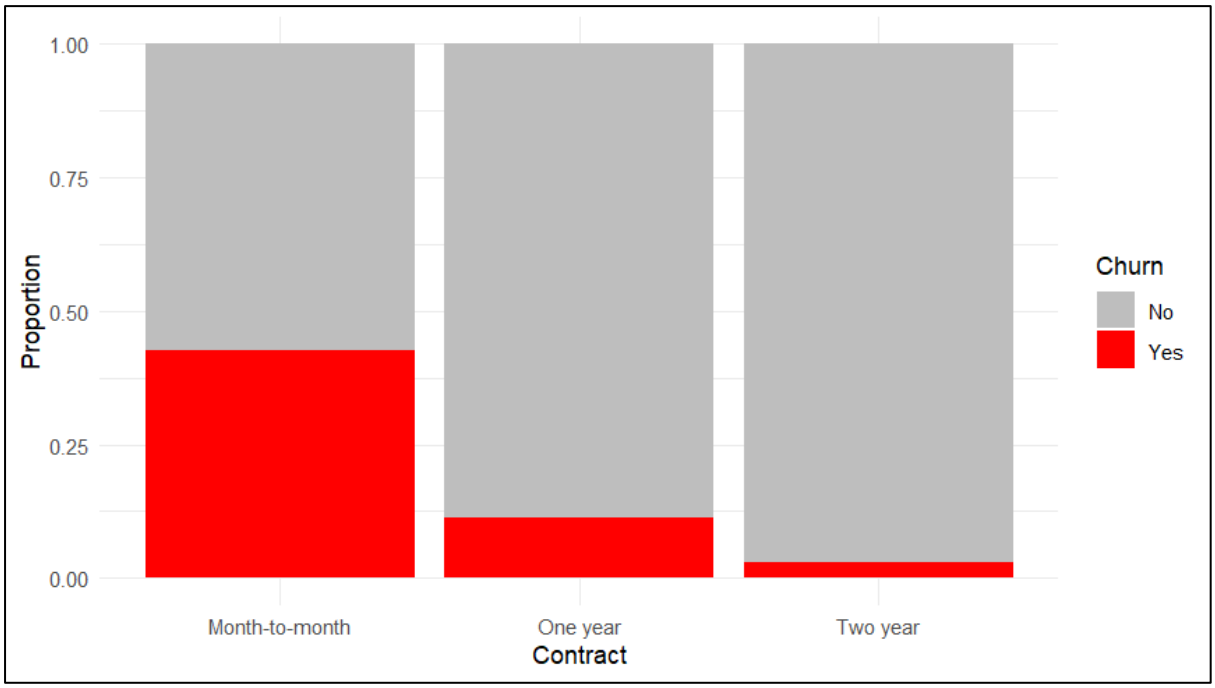

Figure 1: Churn Rate by Contract Type

- The red portion indicates customers who left. Note the significantly higher churn rate among Month-to-Month users compared to those on 1-Year or 2-Year contracts.

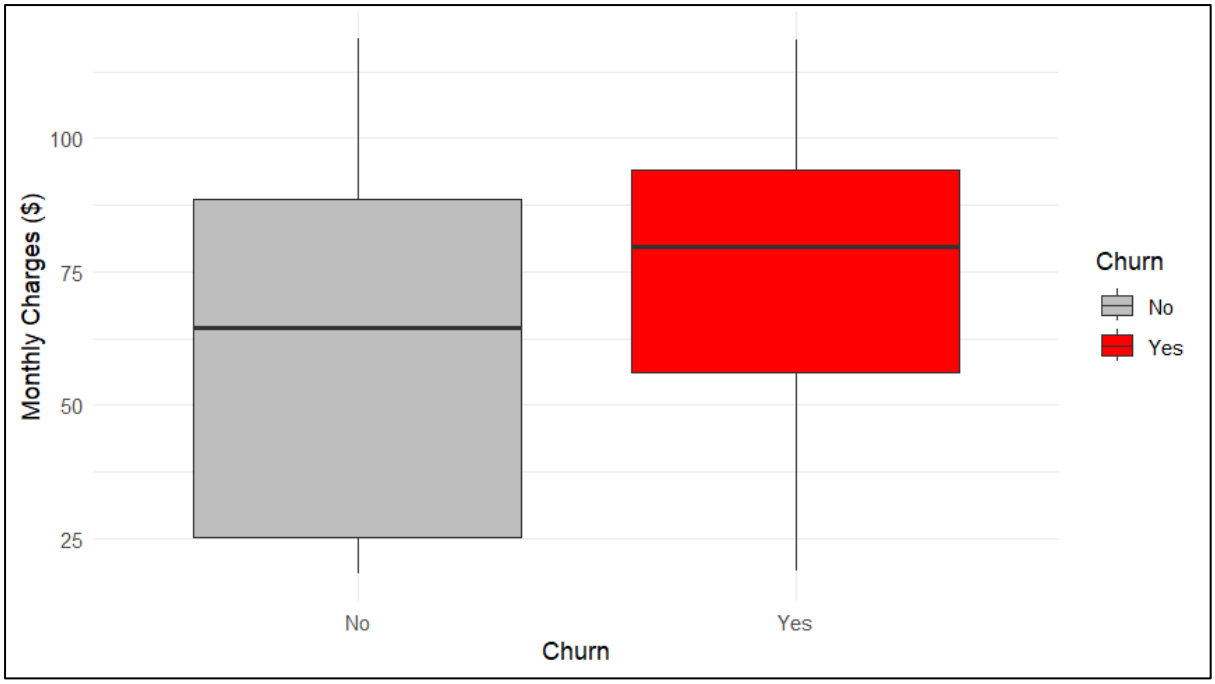

Figure 2.1: Tenure Distribution (The Boxplot of Charges)

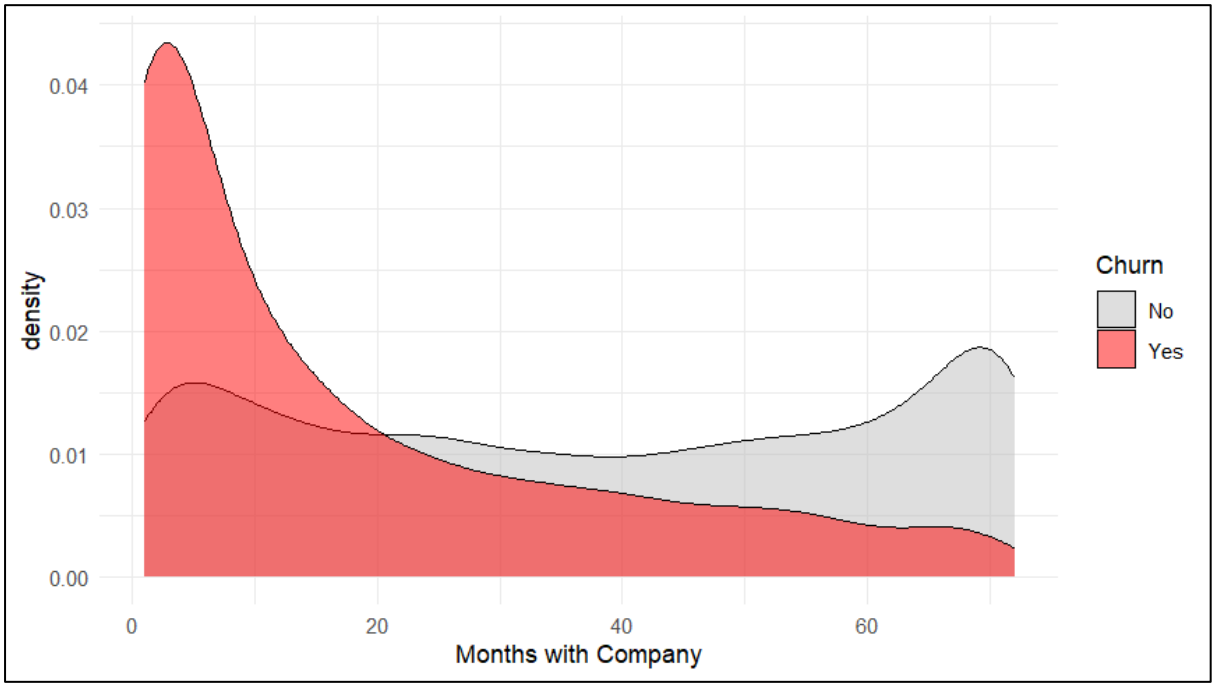

Figure 2.2: Tenure Distribution (Density Plot of Tenure)

- The high density of churn (red area) in the 0–12 month range indicates that new customers are the most vulnerable to leaving.

3. Machine Learning Model

3.1 Algorithm Selection We selected the **Random Forest** algorithm for this classification task. Random Forest is an ensemble learning method that constructs multiple decision trees and merges them to obtain a more accurate and stable prediction. It is particularly effective for handling high-dimensional data and preventing overfitting.

3.2 Model Training

- **Training Set:** 5,275 customers (75%)
- **Test Set:** 1,757 customers (25%)
- **Cross-Validation:** 5-Fold Cross-Validation was applied to ensure the model's reliability.

3.3 Performance Metrics The model was evaluated on the unseen Test Set.

Metric	Score	Interpretation
Accuracy	78.60%	The model correctly predicted customer behavior 79% of the time.
Kappa	0.42	Indicates moderate predictive power beyond random chance.
Sensitivity	0.88	The model is highly effective at identifying loyal customers (True Negatives).

3.4 Confusion Matrix The table below details the model's specific predictions vs. actual outcomes:

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1143  229
       Yes  147  238

               Accuracy : 0.786
                 95% CI : (0.7661, 0.805)
    No Information Rate : 0.7342
    P-Value [Acc > NIR] : 2.934e-07

                  Kappa : 0.4192

 Mcnemar's Test P-Value : 2.950e-05

            Sensitivity : 0.8860
            Specificity : 0.5096
         Pos Pred Value : 0.8331
         Neg Pred Value : 0.6182
             Prevalence : 0.7342
         Detection Rate : 0.6505
   Detection Prevalence : 0.7809
      Balanced Accuracy : 0.6978

       'Positive' Class : No
```

Confusion Matrix showing the count of

- True Positives
- True Negatives
- False Positives
- False Negatives

4. Feature Importance & Recommendations

To understand *why* the model made its decisions, we extracted the "Variable Importance" scores from the Random Forest algorithm.

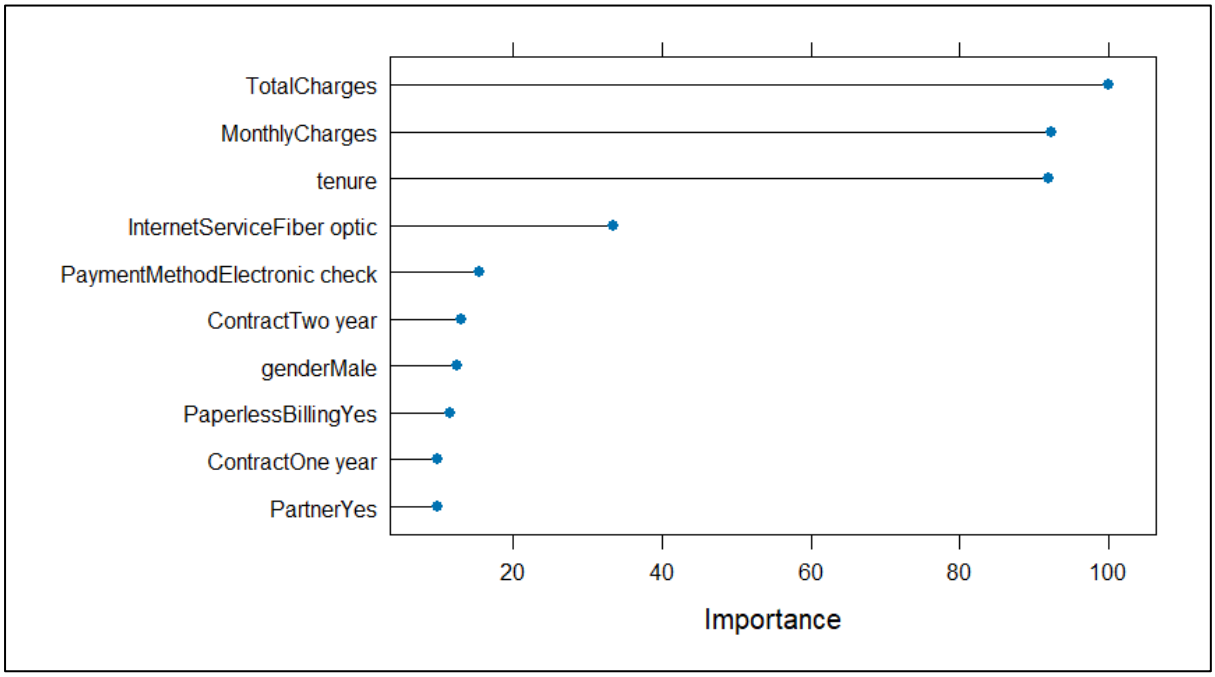

Figure 4: Top 10 Predictors of Churn

- The model identified Contract, Tenure, and Total Charges as the most significant factors influencing customer decisions.

5. Strategic Recommendations

Based on the data, we propose the following actions:

1. **Incentivize Long-Term Contracts:** Since Month-to-Month contracts are the #1 driver of churn, marketing should offer a 5-10% discount or free upgrade to customers who switch to a 1-Year Contract.
2. **The "First Year" Safety Net:** Data shows churn peaks in months 0-12. Implement a specialized support team for new customers to ensure their setup is smooth and satisfaction is high during this critical window.
3. **High-Risk Alerts:** Deploy this Random Forest model into the CRM system. When the model flags a current customer as "High Risk" (Churn Probability > 0.7), automatically send a retention offer (e.g., "One month free") to prevent them from leaving.

6. Appendix A: R Source Code

The following script was used to perform data cleaning, exploratory analysis, and machine learning modeling.

```
# --- PHASE 1: LIBRARY & DATA LOADING ---
library(tidyverse)
library(caret)
library(corrplot)
library(randomForest)

# Load dataset
df <- read.csv("data/churn_data.csv")

# --- PHASE 2: DATA PREPROCESSING ---
# Remove ID, handle missing values, and convert text to factors
clean_df <- df %>%
  select(-customerID) %>%
  na.omit() %>%
  mutate(SeniorCitizen = as.factor(SeniorCitizen)) %>%
  mutate_if(is.character, as.factor)

# Split Data: 75% Training / 25% Testing
set.seed(123)
index <- createDataPartition(clean_df$Churn, p = 0.75, list = FALSE)
train_data <- clean_df[index, ]
test_data  <- clean_df[-index, ]

# --- PHASE 3: EXPLORATORY DATA ANALYSIS (EDA) ---
# 1. Churn by Contract Type
p1 <- ggplot(clean_df, aes(x = Contract, fill = Churn)) +
  geom_bar(position = "fill") +
  labs(title = "Churn Rate by Contract Type", y = "Proportion") +
  theme_minimal()
print(p1)

# 2. Tenure Distribution
p2 <- ggplot(clean_df, aes(x = tenure, fill = Churn)) +
  geom_density(alpha = 0.5) +
  labs(title = "Tenure Density (New vs Loyal Customers)", x = "Months") +
  theme_minimal()
print(p2)

# --- PHASE 4: MACHINE LEARNING (RANDOM FOREST) ---
# Setup Cross-Validation (5-Fold)
ctrl <- trainControl(method = "cv", number = 5)

# Train Model
rf_model <- train(Churn ~ .,
                  data = train_data,
                  method = "rf",
                  trControl = ctrl,
```

```
                  ntree = 100)

print(rf_model)

# --- PHASE 5: EVALUATION & FEATURE IMPORTANCE ---
# Predictions on Test Set
predictions <- predict(rf_model, test_data)

# Confusion Matrix
conf_matrix <- confusionMatrix(predictions, test_data$Churn)
print(conf_matrix)

# Feature Importance
importance <- varImp(rf_model)
plot(importance, top = 10, main = "Top 10 Drivers of Churn")
```