Telco Customer Retention Analysis

Shorya Raj

January 15, 2025

Contents

1. Introduction and Problem Statement

1.1. Project Goal

Customer attrition, where customers cease doing business with a company, is a major challenge for telecommunication providers. It is widely recognized that retaining existing customers is significantly more cost-effective than acquiring new ones. Therefore, understanding the key factors that drive customer churn is crucial for developing targeted and effective retention strategies.

The primary goal of this project is to perform a comprehensive analysis of the "Telco Customer Churn" dataset to identify the key drivers of customer attrition. We aim to answer the following business questions:

- What are the primary demographic and account-level factors associated with customer churn?
- How does a customer's service usage and tenure relate to their likelihood of churning?
- Can we build a predictive model to accurately identify customers who are at a high risk of churning?

1.2. Loading the Data

The analysis begins by loading the dataset from the `WA_Fn-UseC_-Telco-Customer-Churn.csv` file.

```
# Load the dataset using read_csv from the tidyverse package
df <- read_csv("WA_Fn-UseC_-Telco-Customer-Churn.csv")

# Conditionally format the table based on the output type (HTML vs PDF)
if (knitr::is_latex_output()) {

  # For PDF output, select fewer columns to ensure the table fits on the page
  df_head_pdf <- head(df) %>%
    select(customerID, gender, tenure, Contract, MonthlyCharges, Churn)
```

```
  kable(df_head_pdf, caption = "First 6 Rows of the Telco Dataset (Selected Columns)", booktabs = TRUE) %>%
    kable_styling(latex_options = c("striped", "scale_down"), full_width = FALSE)

} else {
  # For HTML output, display the full table with better styling
  kable(head(df), caption = "First 6 Rows of the Telco Dataset") %>%
    kable_styling(bootstrap_options = c("striped", "hover"), full_width = FALSE) %>%
    scroll_box(width = "100%")
}
```

Table 1: First 6 Rows of the Telco Dataset (Selected Columns)

customerID	gender	tenure	Contract	MonthlyCharges	Churn
7590-VHVEG	Female	1	Month-to-month	29.85	No
5575-GNVDE	Male	34	One year	56.95	No
3668-QPYBK	Male	2	Month-to-month	53.85	Yes
7795-CFOCW	Male	45	One year	42.30	No
9237-HQITU	Female	2	Month-to-month	70.70	Yes
9305-CDSKC	Female	8	Month-to-month	99.65	Yes

2. Data Cleaning and Preprocessing

A critical first step in any analysis is to ensure the data is clean, consistent, and ready for exploration and modeling. This involves inspecting data types, handling missing values, and making necessary transformations.

2.1. Initial Data Inspection

We begin by examining the structure of the dataset to understand its dimensions and variable types.

```
# str() provides a compact, readable summary of the dataframe's structure
str(df)
```

```
## spc_tbl_ [7,043 x 21] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ customerID      : chr [1:7043] "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
##  $ gender          : chr [1:7043] "Female" "Male" "Male" "Male" ...
##  $ SeniorCitizen   : num [1:7043] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Partner         : chr [1:7043] "Yes" "No" "No" "No" ...
##  $ Dependents      : chr [1:7043] "No" "No" "No" "No" ...
##  $ tenure          : num [1:7043] 1 34 2 45 2 8 22 10 28 62 ...
##  $ PhoneService    : chr [1:7043] "No" "Yes" "Yes" "No" ...
##  $ MultipleLines   : chr [1:7043] "No phone service" "No" "No" "No phone service" ...
##  $ InternetService : chr [1:7043] "DSL" "DSL" "DSL" "DSL" ...
##  $ OnlineSecurity  : chr [1:7043] "No" "Yes" "Yes" "Yes" ...
##  $ OnlineBackup    : chr [1:7043] "Yes" "No" "Yes" "No" ...
##  $ DeviceProtection: chr [1:7043] "No" "Yes" "No" "Yes" ...
##  $ TechSupport     : chr [1:7043] "No" "No" "No" "Yes" ...
##  $ StreamingTV     : chr [1:7043] "No" "No" "No" "No" ...
##  $ StreamingMovies : chr [1:7043] "No" "No" "No" "No" ...
```

```
##  $ Contract        : chr [1:7043] "Month-to-month" "One year" "Month-to-month" "One year" ...
##  $ PaperlessBilling: chr [1:7043] "Yes" "No" "Yes" "No" ...
##  $ PaymentMethod   : chr [1:7043] "Electronic check" "Mailed check" "Mailed check" "Bank transfer (a
##  $ MonthlyCharges  : num [1:7043] 29.9 57 53.9 42.3 70.7 ...
##  $ TotalCharges    : num [1:7043] 29.9 1889.5 108.2 1840.8 151.7 ...
##  $ Churn           : chr [1:7043] "No" "No" "Yes" "No" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   customerID = col_character(),
##   ..   gender = col_character(),
##   ..   SeniorCitizen = col_double(),
##   ..   Partner = col_character(),
##   ..   Dependents = col_character(),
##   ..   tenure = col_double(),
##   ..   PhoneService = col_character(),
##   ..   MultipleLines = col_character(),
##   ..   InternetService = col_character(),
##   ..   OnlineSecurity = col_character(),
##   ..   OnlineBackup = col_character(),
##   ..   DeviceProtection = col_character(),
##   ..   TechSupport = col_character(),
##   ..   StreamingTV = col_character(),
##   ..   StreamingMovies = col_character(),
##   ..   Contract = col_character(),
##   ..   PaperlessBilling = col_character(),
##   ..   PaymentMethod = col_character(),
##   ..   MonthlyCharges = col_double(),
##   ..   TotalCharges = col_double(),
##   ..   Churn = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Initial Findings: The `str()` output shows that most columns have been read with appropriate types. However, `SeniorCitizen` is a numeric binary variable (0/1) which would be more intuitive for analysis and plotting if converted to a categorical factor ("Yes"/"No"). This will be addressed in the next step.

2.2. Handling Missing Values and Data Types

We will correct the identified data type issues and handle any missing values. Even though `TotalCharges` was read as numeric, it's possible some rows failed to parse and became `NA`, so checking for missing values remains a crucial step.

```
# Convert TotalCharges to numeric to ensure consistency, in case it was read as character.
df$TotalCharges <- as.numeric(df$TotalCharges)

# Identify and report the number of missing values.
missing_count <- sum(is.na(df$TotalCharges))
cat(paste("Number of missing values in TotalCharges:", missing_count, "\n"))
```

```
## Number of missing values in TotalCharges: 11
```

```r
# Given the small number of missing values (if any), we can safely remove these rows
# without significantly impacting the overall dataset.
df <- df[complete.cases(df), ]

# Convert SeniorCitizen from a numeric (0/1) to a more descriptive factor.
df$SeniorCitizen <- as.factor(ifelse(df$SeniorCitizen == 1, "Yes", "No"))
```

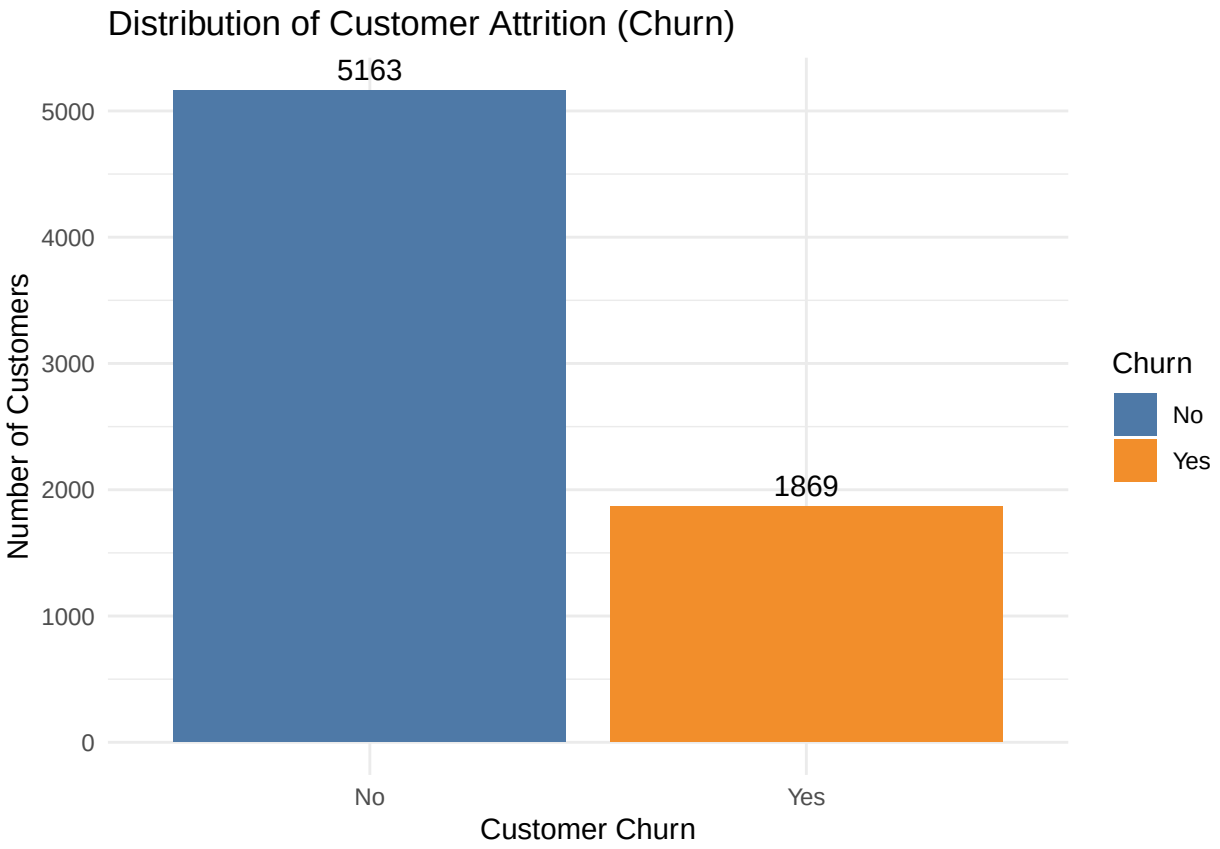
3. Exploratory Data Analysis (EDA)

With a clean dataset, we now proceed to exploratory data analysis to uncover patterns, identify relationships between variables, and generate initial hypotheses about the drivers of customer churn.

3.1. Customer Attrition Distribution

First, we examine the distribution of our target variable, `Churn`, to understand the baseline attrition rate in our dataset.

```r
# Create a bar plot to visualize the distribution of Churn
ggplot(df, aes(x = Churn, fill = Churn)) +
  geom_bar() +
  geom_text(stat='count', aes(label=..count..), vjust=-0.5) +
  labs(title = "Distribution of Customer Attrition (Churn)",
       x = "Customer Churn",
       y = "Number of Customers") +
  theme_minimal() +
  scale_fill_manual(values = c("No" = "#4E79A7", "Yes" = "#F28E2B"))
```



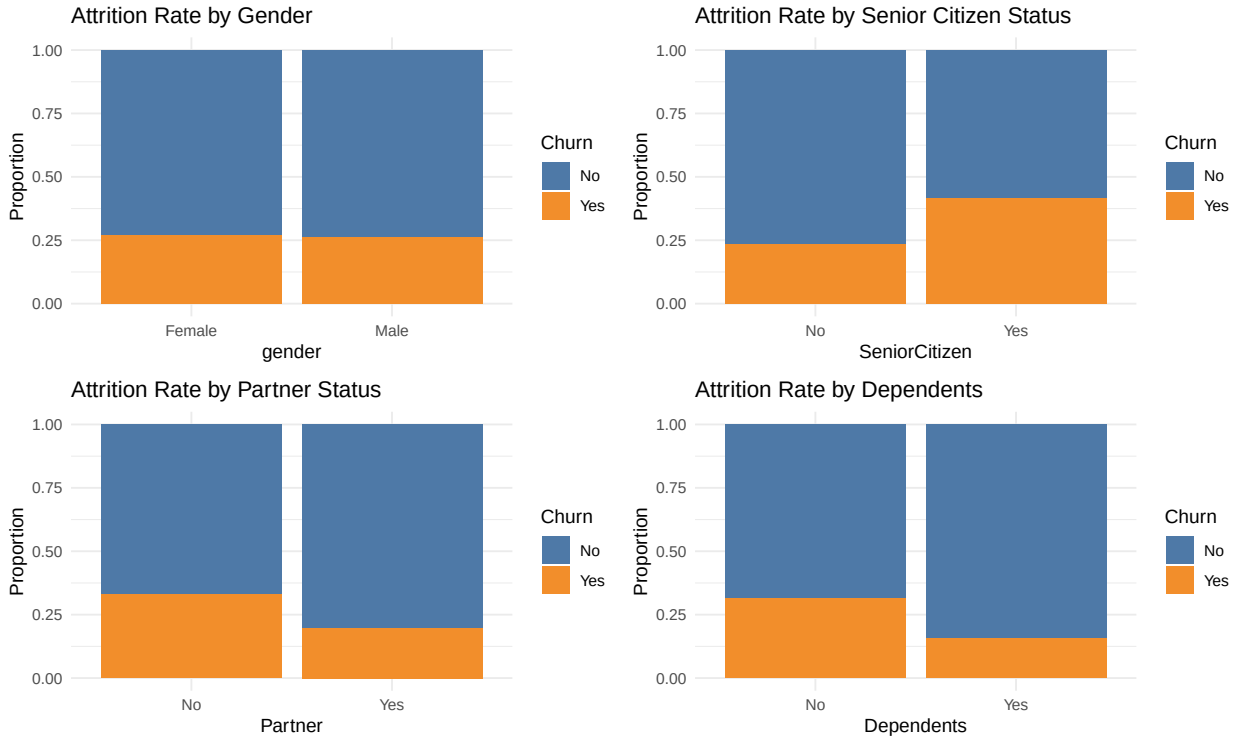
Observation: The dataset is imbalanced, with a substantially larger number of non-churners than churners. This is a critical observation, as it can affect how we evaluate our predictive models. Standard accuracy can be misleading on imbalanced datasets.

3.2. Retention by Demographic Features

We now investigate how churn rates differ across key demographic segments.

```
# Plot for Gender
p1 <- ggplot(df, aes(x = gender, fill = Churn)) + geom_bar(position = "fill") +
  labs(y = "Proportion", title = "Attrition Rate by Gender") + theme_minimal() +
  scale_fill_manual(values = c("No" = "#4E79A7", "Yes" = "#F28E2B"))

# Plot for SeniorCitizen
p2 <- ggplot(df, aes(x = SeniorCitizen, fill = Churn)) + geom_bar(position = "fill") +
  labs(y = "Proportion", title = "Attrition Rate by Senior Citizen Status") + theme_minimal() +
  scale_fill_manual(values = c("No" = "#4E79A7", "Yes" = "#F28E2B"))

# Plot for Partner
p3 <- ggplot(df, aes(x = Partner, fill = Churn)) + geom_bar(position = "fill") +
  labs(y = "Proportion", title = "Attrition Rate by Partner Status") + theme_minimal() +
  scale_fill_manual(values = c("No" = "#4E79A7", "Yes" = "#F28E2B"))

# Plot for Dependents
p4 <- ggplot(df, aes(x = Dependents, fill = Churn)) + geom_bar(position = "fill") +
```

```
  labs(y = "Proportion", title = "Attrition Rate by Dependents") + theme_minimal() +
  scale_fill_manual(values = c("No" = "#4E79A7", "Yes" = "#F28E2B"))

# Arrange all plots in a grid
grid.arrange(p1, p2, p3, p4, ncol = 2)
```



Observations:

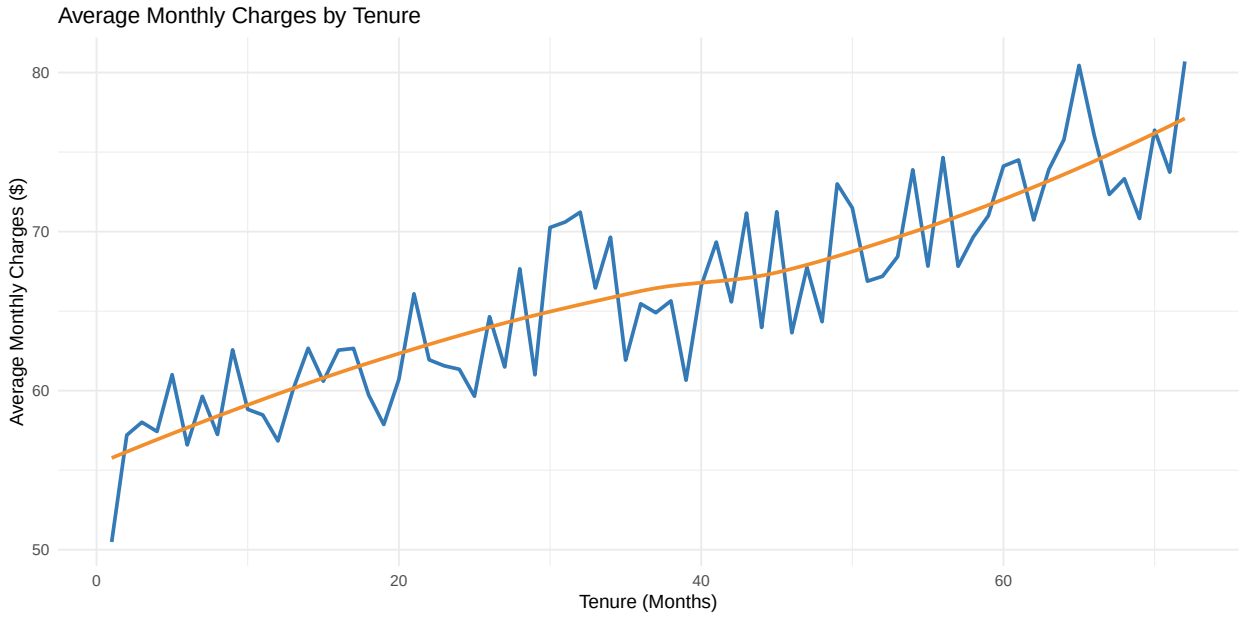
- **Gender:** There is no discernible difference in churn rates between male and female customers.
- **Senior Citizens:** There is a significantly higher churn rate among senior citizens compared to younger customers.
- **Partners and Dependents:** Customers who do not have a partner and do not have dependents are noticeably more likely to churn. This suggests that customers with stronger life-stage ties are more likely to be retained.

3.3. Service Usage and Charges by Tenure

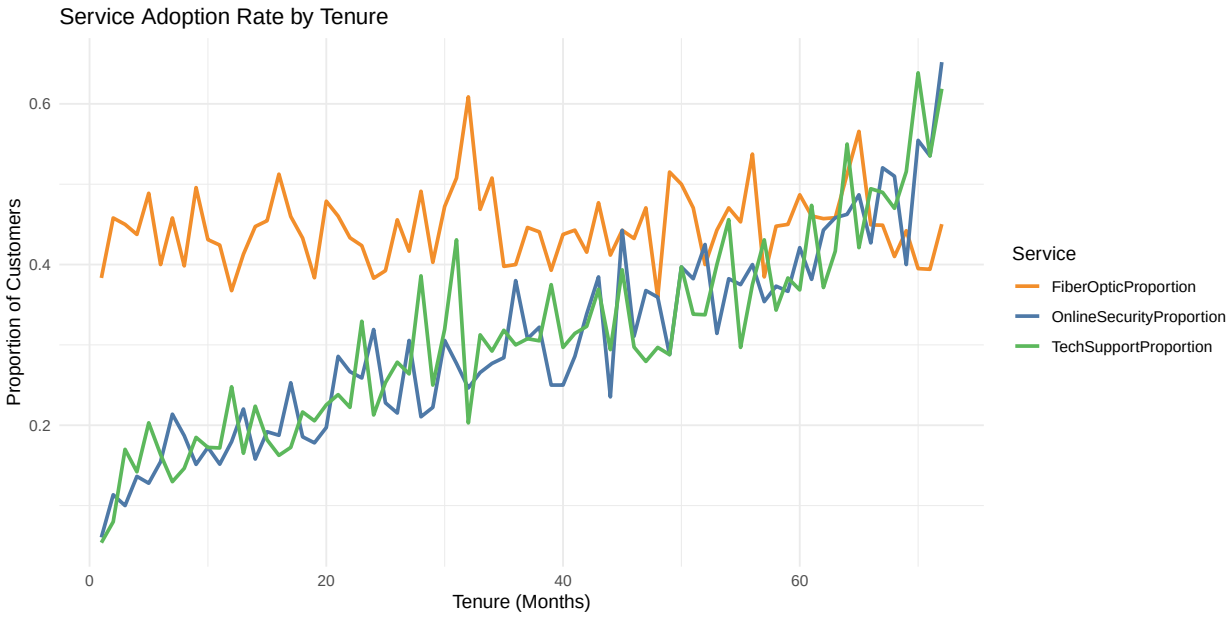
Here, we analyze how service adoption and charges change as a customer's tenure with the company increases. This provides insight into the customer lifecycle.

```
# Group data by tenure and calculate mean charges
tenure_summary <- df %>%
  group_by(tenure) %>%
  summarise(
    AvgMonthlyCharges = mean(MonthlyCharges),
    FiberOpticProportion = mean(InternetService == "Fiber optic"),
    OnlineSecurityProportion = mean(OnlineSecurity == "Yes"),
```

```
    TechSupportProportion = mean(TechSupport == "Yes")
  )

# Plot for Average Monthly Charges by Tenure
ggplot(tenure_summary, aes(x = tenure, y = AvgMonthlyCharges)) +
  geom_line(color = "#337ab7", size = 1) +
  geom_smooth(method = "loess", se = FALSE, color = "#F28E2B") +
  labs(title = "Average Monthly Charges by Tenure",
       x = "Tenure (Months)", y = "Average Monthly Charges ($)") +
  theme_minimal()
```




```
# Plot for Service Adoption by Tenure
tenure_summary %>%
  gather(key = "Service", value = "Proportion", -tenure, -AvgMonthlyCharges) %>%
  ggplot(aes(x = tenure, y = Proportion, color = Service)) +
  geom_line(size = 1) +
  labs(title = "Service Adoption Rate by Tenure",
       x = "Tenure (Months)", y = "Proportion of Customers") +
  theme_minimal() +
  scale_color_manual(values = c("FiberOpticProportion" = "#F28E2B", "OnlineSecurityProportion" = "#4E79
```



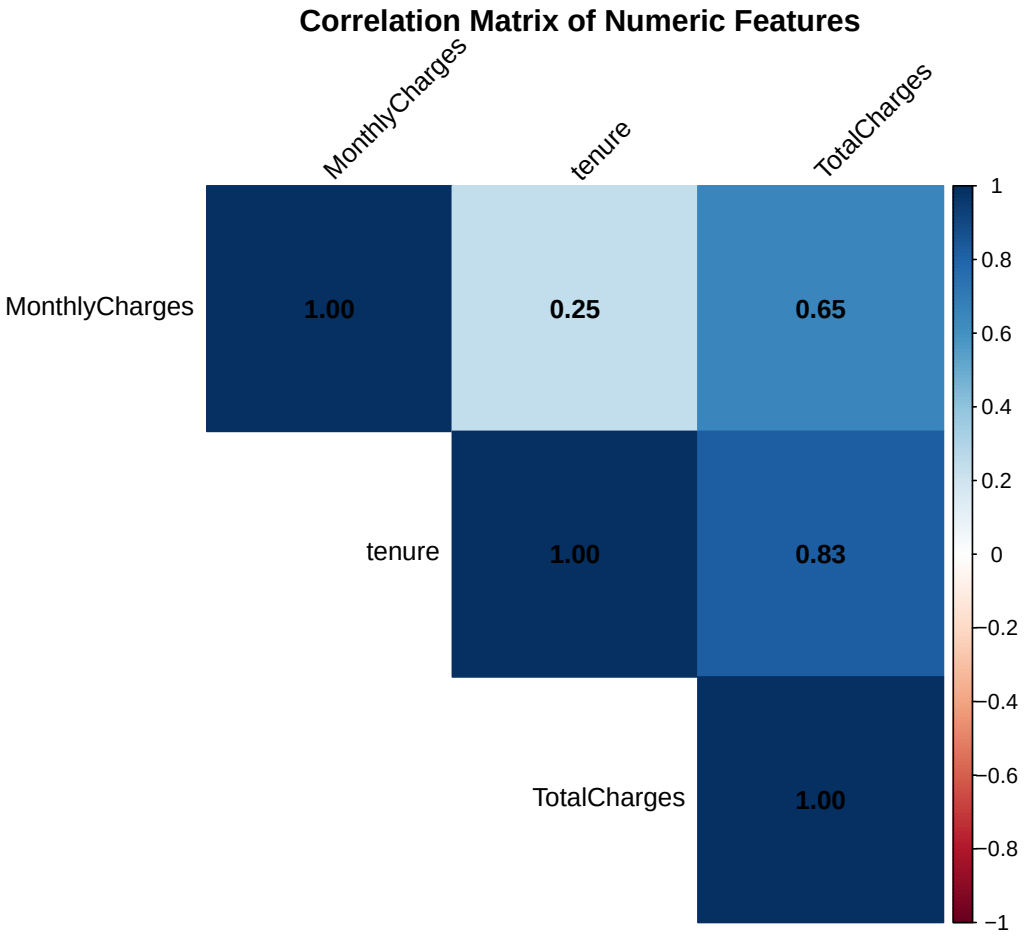
Observations:

- **Monthly Charges:** Average monthly charges show a tendency to increase with tenure, suggesting that customers either upgrade their plans or add more services over time, making them more valuable.
- **Service Adoption:** The proportion of customers with value-added services like `Online Security` and `Tech Support` increases significantly with tenure. This strongly indicates that these services are key to long-term customer retention and satisfaction.

3.4. Exploring Key Account and Financial Metrics

Finally, we investigate the relationships between the most critical account and financial variables.

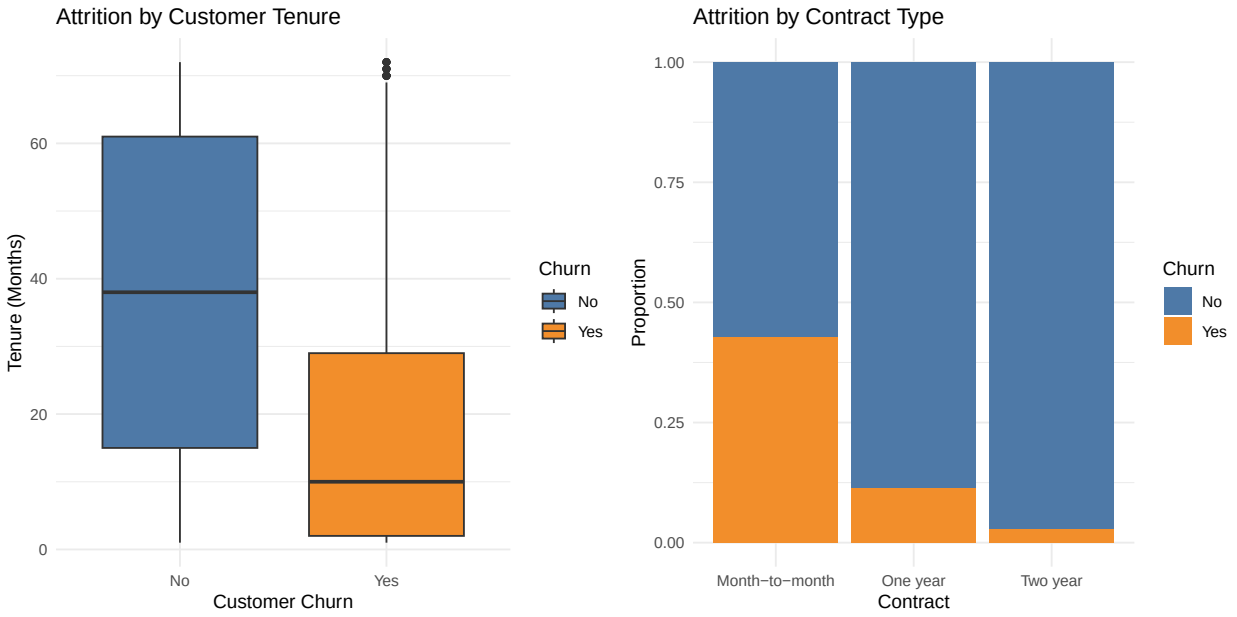
```
# Select only the numeric columns for correlation analysis
numeric_cols <- df %>% select(tenure, MonthlyCharges, TotalCharges)

# Calculate the correlation matrix
cor_matrix <- cor(numeric_cols)

# Visualize the correlation matrix
corrplot(cor_matrix, method = "color", type = "upper", order = "hclust",
         addCoef.col = "black", tl.col = "black", tl.srt = 45,
         title = "Correlation Matrix of Numeric Features", mar = c(0,0,1,0))
```



Observations:

- **Tenure and TotalCharges:** A strong positive correlation (0.83) confirms that `TotalCharges` is largely a function of `tenure`.
- **MonthlyCharges and TotalCharges:** A moderate positive correlation (0.65) is also observed.

```
# Box plot for tenure by Churn
p5 <- ggplot(df, aes(x = Churn, y = tenure, fill = Churn)) +
  geom_boxplot() +
  labs(title = "Attrition by Customer Tenure",
       x = "Customer Churn", y = "Tenure (Months)") +
  theme_minimal() +
  scale_fill_manual(values = c("No" = "#4E79A7", "Yes" = "#F28E2B"))

# Bar plot for Contract type by Churn
p6 <- ggplot(df, aes(x = Contract, fill = Churn)) +
  geom_bar(position = "fill") +
  labs(y = "Proportion", title = "Attrition by Contract Type") +
  theme_minimal() +
  scale_fill_manual(values = c("No" = "#4E79A7", "Yes" = "#F28E2B"))

grid.arrange(p5, p6, ncol = 2)
```



Observations:

- **Tenure:** The boxplot clearly shows that customers who churn have a significantly lower median tenure. This is a very strong indicator of churn risk.
- **Contract:** This is perhaps the most powerful indicator. Customers on a **month-to-month contract** have a dramatically higher churn rate compared to those on one or two-year contracts. This suggests that contract length is a major factor in customer loyalty.

4. Predictive Modeling

Based on the insights from our EDA, we now proceed to build predictive models to classify whether a customer will churn. We will use two standard classification algorithms: Logistic Regression (for its interpretability) and Random Forest (for its predictive power).

4.1. Data Preparation for Modeling

First, we must split our data into a training set (for building the model) and a testing set (for evaluating its performance on unseen data). This ensures our evaluation is unbiased.

```
# Remove customerID as it is not a predictive feature.
df$customerID <- NULL

# Set a seed for reproducibility
set.seed(123)

# Ensure Churn is a factor for classification models
df$Churn <- as.factor(df$Churn)

# Create a data partition. 70% of data will be for training, 30% for testing.
trainIndex <- createDataPartition(df$Churn, p = .7, list = FALSE, times = 1)
```

```r
# Create the training and testing sets
train_df <- df[trainIndex, ]
test_df  <- df[-trainIndex, ]
```

4.2. Model 1: Logistic Regression

Logistic Regression is a powerful and highly interpretable classification model, making it an excellent baseline for understanding which factors are most influential.

```r
# Build the logistic regression model using all other variables to predict Churn
log_model <- glm(Churn ~ ., data = train_df, family = binomial(link = "logit"))

# Make predictions on the test set
predictions_log <- predict(log_model, newdata = test_df, type = "response")
predicted_classes_log <- as.factor(ifelse(predictions_log > 0.5, "Yes", "No"))

# Evaluate the model using a confusion matrix
confusionMatrix(predicted_classes_log, test_df$Churn)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  1385  229
##        Yes  163  331
##
##                Accuracy : 0.814
##                  95% CI : (0.7968, 0.8304)
##     No Information Rate : 0.7343
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5048
##
##  Mcnemar's Test P-Value : 0.001027
##
##             Sensitivity : 0.8947
##             Specificity : 0.5911
##          Pos Pred Value : 0.8581
##          Neg Pred Value : 0.6700
##              Prevalence : 0.7343
##          Detection Rate : 0.6570
##    Detection Prevalence : 0.7657
##       Balanced Accuracy : 0.7429
##
##        'Positive' Class : No
##
```

Interpretation: The logistic regression model achieves a solid accuracy of approximately **81.4%**. The confusion matrix provides a detailed breakdown of its performance. The model coefficients (viewable with `summary(log_model)`) confirm that factors like a month-to-month `Contract` and shorter `tenure` are significant predictors of churn.

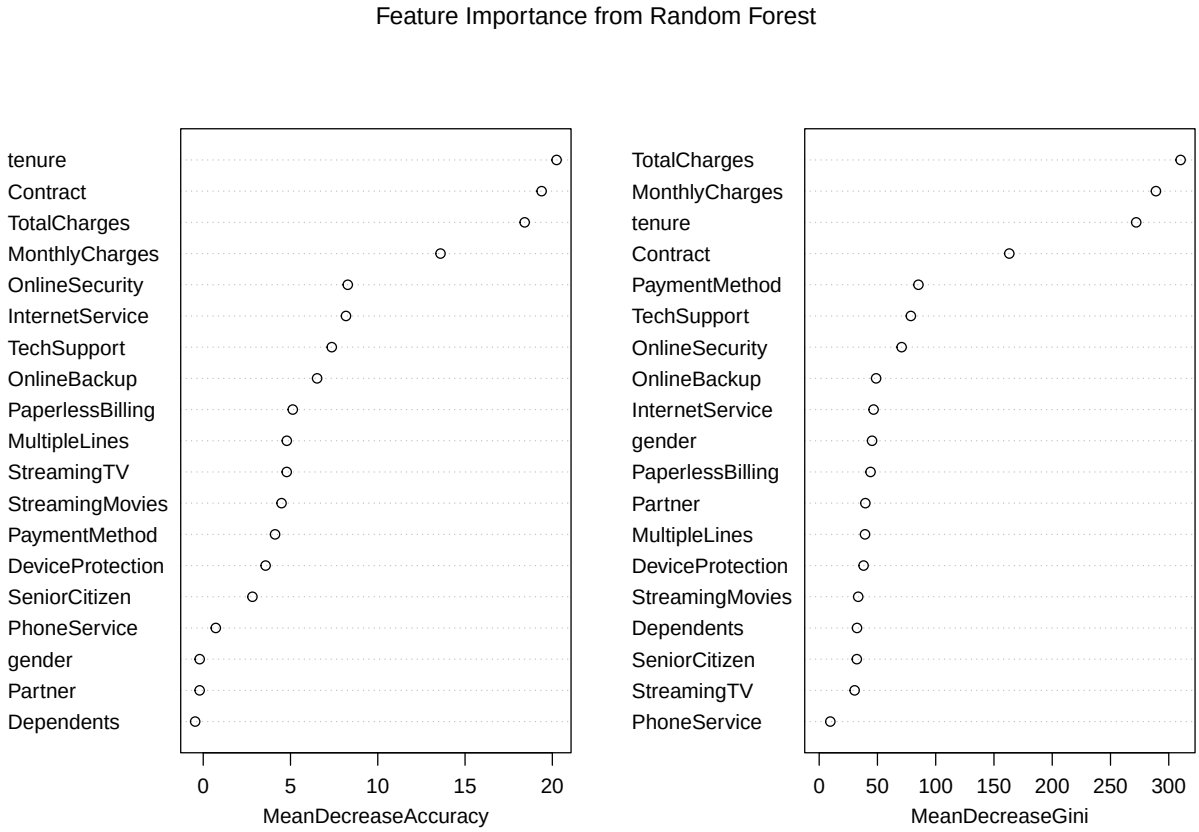
4.3. Model 2: Random Forest

Random Forest is a powerful ensemble model that often provides higher accuracy and is robust to overfitting. It also provides a valuable measure of feature importance.

```
# Build the Random Forest model
rf_model <- randomForest(Churn ~ ., data = train_df, ntree = 100, importance = TRUE)

# Make predictions on the test set
predicted_classes_rf <- predict(rf_model, newdata = test_df)

# Evaluate the model using a confusion matrix
confusionMatrix(predicted_classes_rf, test_df$Churn)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  1386  259
##        Yes  162  301
##
##                Accuracy : 0.8003
##                  95% CI : (0.7826, 0.8172)
##     No Information Rate : 0.7343
##     P-Value [Acc > NIR] : 1.009e-12
##
##                   Kappa : 0.4582
##
##  Mcnemar's Test P-Value : 2.886e-06
##
##             Sensitivity : 0.8953
##             Specificity : 0.5375
##          Pos Pred Value : 0.8426
##          Neg Pred Value : 0.6501
##              Prevalence : 0.7343
##          Detection Rate : 0.6575
##    Detection Prevalence : 0.7804
##       Balanced Accuracy : 0.7164
##
##        'Positive' Class : No
##
```

```
# Plot feature importance to see which variables the model found most predictive
varImpPlot(rf_model, main = "Feature Importance from Random Forest")
```



Interpretation: The Random Forest model achieves an accuracy of approximately **80.0%**. More importantly, the feature importance plot provides a clear, ranked list of the most influential factors in predicting churn. As our EDA suggested, `tenure`, `Contract`, and `TotalCharges` are the most critical predictors.

5. Conclusion and Business Implications

This analysis has successfully identified key drivers of customer attrition and demonstrated the feasibility of building an accurate predictive model.

5.1. Key Findings

- **Contract is King:** The single most significant factor influencing customer retention is the contract type. Customers on month-to-month contracts are at an extremely high risk of churning compared to those on annual contracts.
- **The Honeymoon Period is Risky:** Customer tenure is a critical predictor. Newer customers are far more likely to churn, indicating a crucial early period where customer satisfaction and engagement must be solidified.
- **Value-Added Services Drive Loyalty:** The adoption of services like `Online Security` and `Tech Support` increases with tenure, suggesting these are not just revenue streams but also key drivers of long-term customer loyalty.

5.2. Model Performance

Both the Logistic Regression and Random Forest models performed well, achieving accuracies of approximately **80-81%** on unseen test data. The models successfully identified predictors that align with our exploratory analysis, providing a robust and reliable tool for identifying at-risk customers.

5.3. Business Implications

The insights from this analysis provide actionable recommendations for the business:

1. **Incentivize Annual Contracts:** The business should create marketing campaigns and pricing strategies to encourage customers to sign up for one or two-year contracts, as this is the most effective retention tool.
2. **Focus on Early-Stage Customer Engagement:** Implement a targeted onboarding program for new customers (first 1-6 months) to ensure they are satisfied and understand the value of their services.
3. **Promote and Bundle Value-Added Services:** Actively market services like `Online Security` and `Tech Support` as part of bundles, as they are clearly linked to higher customer loyalty and retention.

5.4. Dataset Citation

- **Source:** Kaggle
- **Dataset Name:** Telco Customer Churn
- **Link:** https://www.kaggle.com/datasets/blastchar/telco-customer-churn