Telecom Customer Churn Prediction Project Proposal

Ashish Yakub Beary

2025-04-04

Research Question

- What factors most significantly predict customer churn in the telecom industry?
- Secondary questions:
  - How do service usage patterns and demographics correlate with churn probability?
  - Which specific services or contract features have the strongest protective effect against churn?
  - Can we identify high-risk customers early to implement targeted retention strategies?

Cases

- Each case represents an individual customer of a telecommunications company
- The dataset contains 7,043 unique customers
- Each record includes comprehensive information about customer demographics, service subscriptions, billing information, and churn status

Method of Data Collection

- The data was collected by the telecommunications company from their customer relationship management (CRM) system and billing databases
- It represents historical customer data from account creation through either current active status or account termination
- The information includes service subscriptions, billing details, demographic information, and geographic data

Type of Study

- This is an observational study as it analyzes existing customer data without experimental manipulation
- The study examines patterns and relationships in historical data to identify predictive factors for customer churn

Data Source

- Primary dataset: telecom_customer_churn.csv (7,043 rows × 38 columns)
- Supplementary datasets:
  - telecom_zipcode_population.csv - population data by zip code
  - telecom_data_dictionary.csv - metadata describing each variable

Variables

Response Variable

- **Customer Status** (categorical): Customer's current status, with "Churned" indicating customers who have left the company
- For binary analysis purposes, this will be transformed into a churn indicator (Yes/No)

Explanatory Variables

Demographics

- Gender (categorical)
- Age (numerical)
- Married (categorical - Yes/No)
- Number of Dependents (numerical)
- City (categorical)
- Zip Code (categorical)
- Geographic location (Latitude/Longitude)

Account Information

- Number of Referrals (numerical)
- Tenure in Months (numerical)
- Offer (categorical)
- Contract (categorical - Month-to-Month, One Year, Two Year)
- Paperless Billing (categorical - Yes/No)
- Payment Method (categorical)

Service Subscriptions

- Phone Service (categorical - Yes/No)
- Multiple Lines (categorical)
- Internet Service (categorical - Yes/No)
- Internet Type (categorical)
- Avg Monthly GB Download (numerical)
- Online Security (categorical - Yes/No)
- Online Backup (categorical - Yes/No)
- Device Protection Plan (categorical - Yes/No)
- Premium Tech Support (categorical - Yes/No)
- Streaming TV (categorical - Yes/No)
- Streaming Movies (categorical - Yes/No)
- Streaming Music (categorical - Yes/No)
- Unlimited Data (categorical - Yes/No)

Financial Metrics

- Monthly Charge (numerical)
- Total Charges (numerical)
- Total Refunds (numerical)
- Total Extra Data Charges (numerical)

- Total Long Distance Charges (numerical)
- Total Revenue (numerical)
- Avg Monthly Long Distance Charges (numerical)

Churn Details (for churned customers only)

- Churn Category (categorical)
- Churn Reason (categorical)

Preliminary Data Analysis

```r
# Load the datasets
telecom_churn <- read.csv("telecom_customer_churn.csv", stringsAsFactors = TRUE)
zipcode_population <- read.csv("telecom_zipcode_population.csv")
data_dictionary <- read.csv("telecom_data_dictionary.csv", encoding = "CP1252")

# Clean column names
telecom_churn <- clean_names(telecom_churn)
zipcode_population <- clean_names(zipcode_population)
data_dictionary <- clean_names(data_dictionary)

# Display dataset structure
str(telecom_churn)
```

```
## 'data.frame':    7043 obs. of  38 variables:
##  $ customer_id                      : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 1 2 3 4 5 6
##  $ gender                           : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 1 2 1 1 ...
##  $ age                              : int  37 46 50 78 75 23 67 52 68 43 ...
##  $ married                          : Factor w/ 2 levels "No","Yes": 2 1 1 2 2 1 2 2 1 2 ...
##  $ number_of_dependents             : int  0 0 0 0 0 3 0 0 0 1 ...
##  $ city                             : Factor w/ 1106 levels "Acampo","Acton",..: 347 369 223 588 140
##  $ zip_code                         : int  93225 91206 92627 94553 93010 95345 93437 94558 93063 9568
##  $ latitude                         : num  34.8 34.2 33.6 38 34.2 ...
##  $ longitude                        : num  -119 -118 -118 -122 -119 ...
##  $ number_of_referrals              : int  2 0 0 1 3 0 1 8 0 3 ...
##  $ tenure_in_months                 : int  9 9 4 13 3 9 71 63 7 65 ...
##  $ offer                            : Factor w/ 6 levels "None","Offer A",..: 1 1 6 5 1 6 2 3 6 1 ..
##  $ phone_service                    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ avg_monthly_long_distance_charges: num  42.39 10.69 33.65 27.82 7.38 ...
##  $ multiple_lines                   : Factor w/ 3 levels "","No","Yes": 2 3 2 2 2 2 2 3 2 3 ...
##  $ internet_service                 : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ internet_type                    : Factor w/ 4 levels "","Cable","DSL",..: 2 2 4 4 4 2 4 4 3 2 ..
##  $ avg_monthly_gb_download          : int  16 10 30 4 11 73 14 7 21 14 ...
##  $ online_security                  : Factor w/ 3 levels "","No","Yes": 2 2 2 2 2 2 3 3 3 3 ...
##  $ online_backup                    : Factor w/ 3 levels "","No","Yes": 3 2 2 3 2 2 3 2 2 3 ...
##  $ device_protection_plan           : Factor w/ 3 levels "","No","Yes": 2 2 3 3 2 2 3 2 2 3 ...
##  $ premium_tech_support             : Factor w/ 3 levels "","No","Yes": 3 2 2 2 3 3 3 3 2 3 ...
##  $ streaming_tv                     : Factor w/ 3 levels "","No","Yes": 3 2 2 3 3 3 3 2 2 3 ...
##  $ streaming_movies                 : Factor w/ 3 levels "","No","Yes": 2 3 2 3 2 3 3 2 2 3 ...
##  $ streaming_music                  : Factor w/ 3 levels "","No","Yes": 2 3 2 2 2 3 3 2 2 3 ...
##  $ unlimited_data                   : Factor w/ 3 levels "","No","Yes": 3 2 3 3 3 3 3 2 3 3 ...
```

```
##  $ contract                        : Factor w/ 3 levels "Month-to-Month",..: 2 1 1 1 1 1 3 3 3 3 ..
##  $ paperless_billing               : Factor w/ 2 levels "No","Yes": 2 1 2 2 2 2 2 2 2 2 ...
##  $ payment_method                  : Factor w/ 3 levels "Bank Withdrawal",..: 2 2 1 1 2 2 1 2 1 2 .
##  $ monthly_charge                  : num  65.6 -4 73.9 98 83.9 ...
##  $ total_charges                   : num  593 542 281 1238 267 ...
##  $ total_refunds                   : num  0 38.3 0 0 0 ...
##  $ total_extra_data_charges        : int  0 10 0 0 0 0 0 20 0 0 ...
##  $ total_long_distance_charges     : num  381.5 96.2 134.6 361.7 22.1 ...
##  $ total_revenue                   : num  975 610 415 1600 290 ...
##  $ customer_status                 : Factor w/ 3 levels "Churned","Joined",..: 3 3 1 1 1 3 3 3 3 3
##  $ churn_category                  : Factor w/ 6 levels "","Attitude",..: 1 1 3 4 4 1 1 1 1 1 ...
##  $ churn_reason                    : Factor w/ 21 levels "","Attitude of service provider",..: 1 1
```

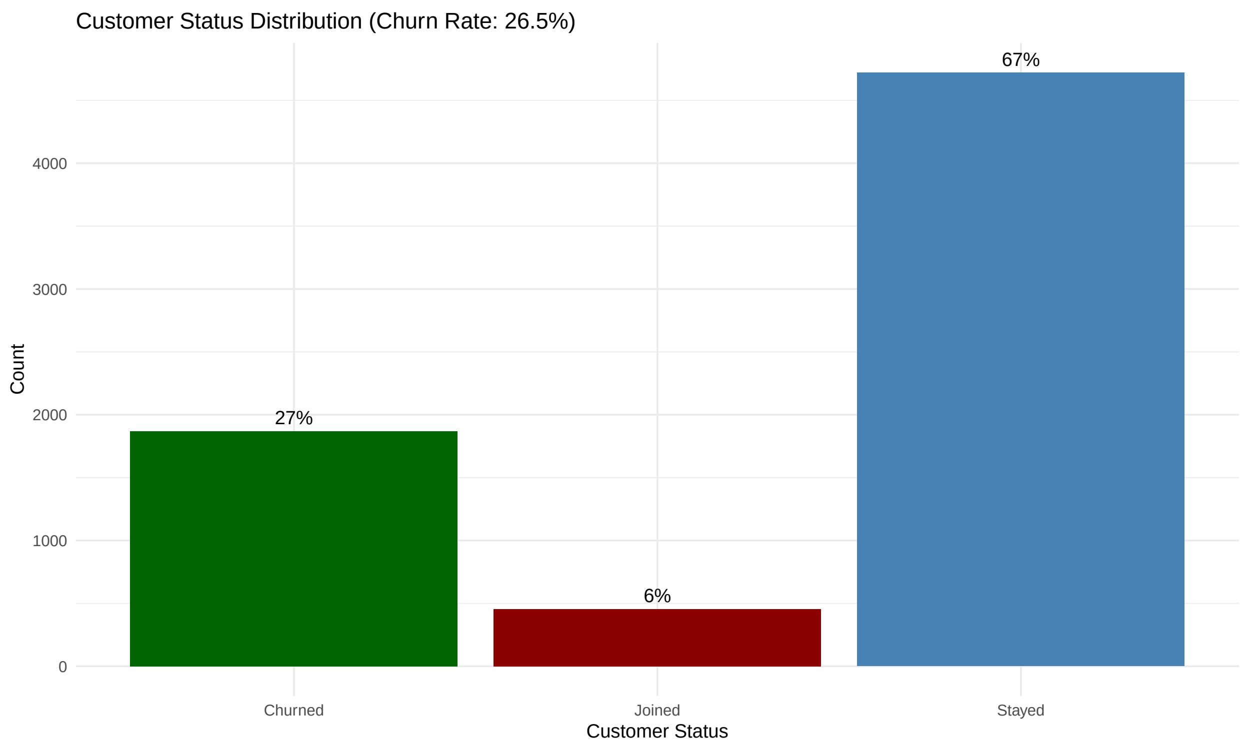
```r
# Create binary churn variable for analysis
telecom_churn$churned <- ifelse(telecom_churn$customer_status == "Churned", "Yes", "No")
telecom_churn$churned <- as.factor(telecom_churn$churned)

# Basic summary statistics
summary(telecom_churn[c("age", "tenure_in_months", "number_of_dependents",
                         "avg_monthly_gb_download", "monthly_charge", "total_charges")])
```

```
##       age        tenure_in_months number_of_dependents avg_monthly_gb_download
##  Min.   :19.00   Min.   : 1.00    Min.   :0.0000       Min.   : 2.00
##  1st Qu.:32.00   1st Qu.: 9.00    1st Qu.:0.0000       1st Qu.:13.00
##  Median :46.00   Median :29.00    Median :0.0000       Median :21.00
##  Mean   :46.51   Mean   :32.39    Mean   :0.4687       Mean   :26.19
##  3rd Qu.:60.00   3rd Qu.:55.00    3rd Qu.:0.0000       3rd Qu.:30.00
##  Max.   :80.00   Max.   :72.00    Max.   :9.0000       Max.   :85.00
##                                                        NA's   :1526
##  monthly_charge   total_charges
##  Min.   :-10.00   Min.   :  18.8
##  1st Qu.: 30.40   1st Qu.: 400.1
##  Median : 70.05   Median :1394.5
##  Mean   : 63.60   Mean   :2280.4
##  3rd Qu.: 89.75   3rd Qu.:3786.6
##  Max.   :118.75   Max.   :8684.8
##
```

```r
# Calculate overall churn rate
churn_rate <- mean(telecom_churn$customer_status == "Churned") * 100

# Visualize the churn distribution
ggplot(telecom_churn, aes(x = customer_status)) +
  geom_bar(fill = c("darkgreen", "darkred", "steelblue")) +
  geom_text(stat = "count", aes(label = scales::percent(..count../sum(after_stat(count)))),
            vjust = -0.5) +
  labs(title = paste0("Customer Status Distribution (Churn Rate: ", round(churn_rate, 1), "%)"),
       x = "Customer Status",
       y = "Count") +
  theme_minimal()
```

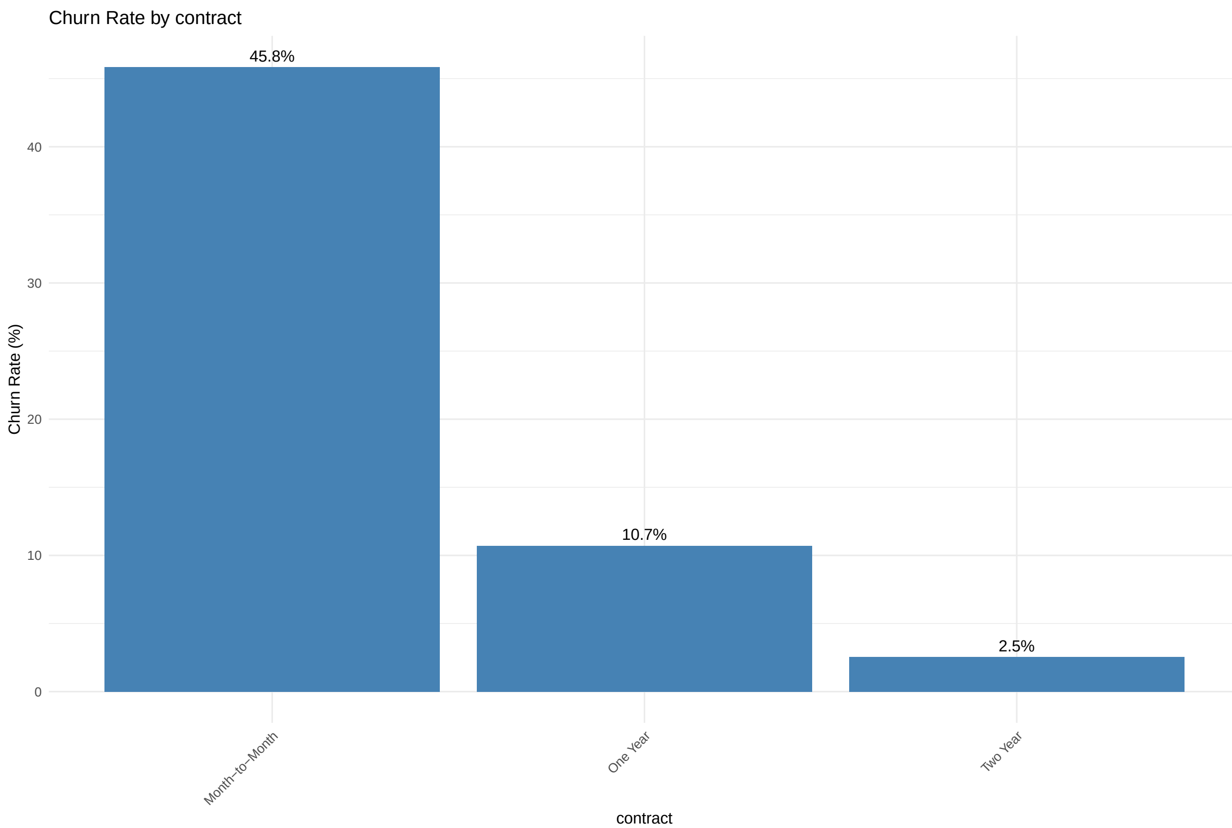
```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
```

```
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```r
# Analyze categorical variables by churn status
cat_vars <- c("contract", "internet_type", "payment_method", "offer", "paperless_billing")

# Create a function to plot churn rate by category
plot_churn_by_category <- function(data, variable) {
  # Calculate percentages
  churn_by_cat <- data %>%
    group_by(!!sym(variable)) %>%
    summarize(
      total = n(),
      churned = sum(customer_status == "Churned"),
      churn_rate = churned / total * 100
    ) %>%
    arrange(desc(churn_rate))

  # Create plot
  ggplot(churn_by_cat, aes(x = reorder(!!sym(variable), -churn_rate), y = churn_rate)) +
    geom_bar(stat = "identity", fill = "steelblue") +
    geom_text(aes(label = paste0(round(churn_rate, 1), "%")), vjust = -0.5) +
    labs(title = paste("Churn Rate by", variable),
         x = variable,
         y = "Churn Rate (%)") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
}
```

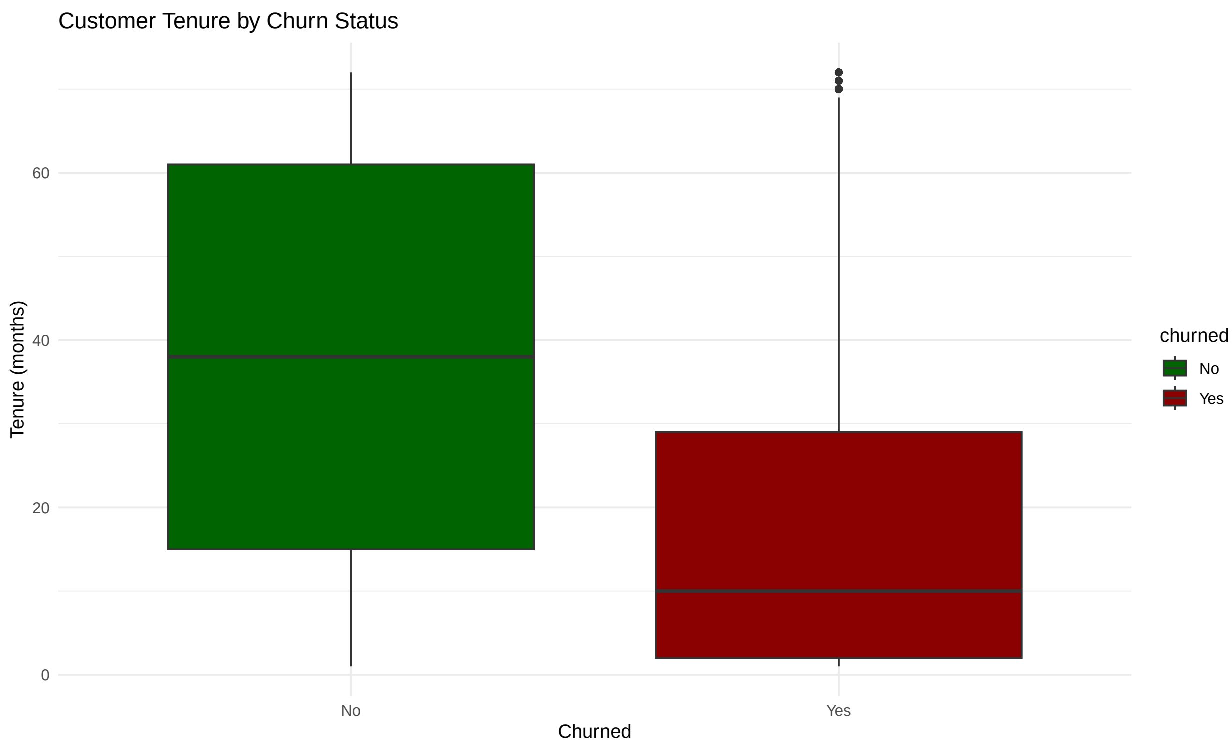
```r
# Plot for contract type
plot_churn_by_category(telecom_churn, "contract")
```




```r
# Analyze numerical variables by churn status
num_vars <- c("age", "tenure_in_months", "monthly_charge", "total_charges", "avg_monthly_gb_download")

# Create summary statistics by churn status
num_summary <- telecom_churn %>%
  group_by(churned) %>%
  summarize(
    avg_age = mean(age, na.rm = TRUE),
    avg_tenure = mean(tenure_in_months, na.rm = TRUE),
    avg_monthly_charge = mean(monthly_charge, na.rm = TRUE),
    avg_total_charges = mean(total_charges, na.rm = TRUE),
    avg_monthly_download = mean(avg_monthly_gb_download, na.rm = TRUE)
  )

print(num_summary)
```

```
## # A tibble: 2 x 6
##   churned avg_age avg_tenure avg_monthly_charge avg_total_charges
##   <fct>     <dbl>      <dbl>              <dbl>             <dbl>
## 1 No         45.3       37.6               60.1             2551.
## 2 Yes        49.7       18.0               73.3             1532.
## # i 1 more variable: avg_monthly_download <dbl>
```

```r
# Visualize tenure by churn status
ggplot(telecom_churn, aes(x = churned, y = tenure_in_months, fill = churned)) +
  geom_boxplot() +
  labs(title = "Customer Tenure by Churn Status",
       x = "Churned",
       y = "Tenure (months)") +
  scale_fill_manual(values = c("darkgreen", "darkred")) +
  theme_minimal()
```



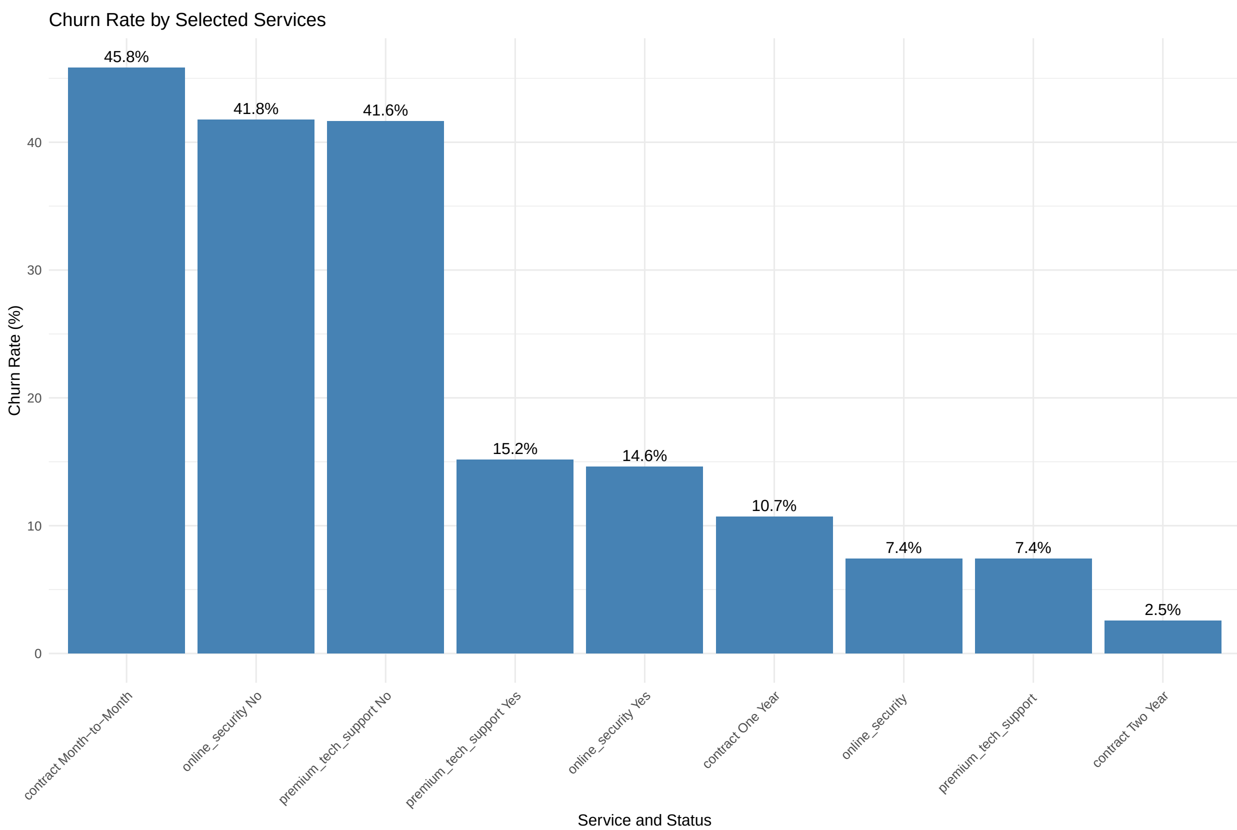

```r
# Analyze service adoption and its impact on churn
service_vars <- c("phone_service", "multiple_lines", "online_security",
                  "online_backup", "device_protection_plan", "premium_tech_support",
                  "streaming_tv", "streaming_movies", "streaming_music", "unlimited_data")

# Specify which services to include in the visualization
services_to_plot <- c("online_security", "premium_tech_support", "contract")

# Create a function to calculate and visualize service impact on churn
service_impact <- function(data, service_var) {
  # Filter out NA values
  data_filtered <- data %>% filter(!is.na(!!sym(service_var)))

  # Calculate churn rates
  service_churn <- data_filtered %>%
    group_by(!!sym(service_var)) %>%
    summarize(
      total = n(),
      churned = sum(customer_status == "Churned"),
      churn_rate = churned / total * 100
```

```r
    )

  return(service_churn)
}

# Example for one service
online_security_impact <- service_impact(telecom_churn, "online_security")
print(online_security_impact)
```

```
## # A tibble: 3 x 4
##   online_security total churned churn_rate
##   <fct>           <int>   <int>      <dbl>
## 1 ""               1526     113       7.40
## 2 "No"             3498    1461      41.8
## 3 "Yes"            2019     295      14.6
```

```r
# Create an empty data frame with the right structure
services_plot_data <- data.frame(
  service = character(),
  service_value = character(),
  total = numeric(),
  churned = numeric(),
  churn_rate = numeric(),
  stringsAsFactors = FALSE
)

# Loop through each service
for (service_name in services_to_plot) {
  # Get the churn data for this service
  service_data <- service_impact(telecom_churn, service_name)

  # Extract the service value column (which has a dynamic name)
  service_values <- service_data[[1]]  # The first column contains the service values

  # Create a new data frame with consistent column names
  temp_df <- data.frame(
    service = service_name,
    service_value = as.character(service_values),
    total = service_data$total,
    churned = service_data$churned,
    churn_rate = service_data$churn_rate,
    stringsAsFactors = FALSE
  )

  # Add to the main data frame
  services_plot_data <- rbind(services_plot_data, temp_df)
}

# Plot with the corrected data structure
ggplot(services_plot_data, aes(x = reorder(paste(service, service_value), -churn_rate), y = churn_rate)
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = paste0(round(churn_rate, 1), "%")), vjust = -0.5) +
  labs(title = "Churn Rate by Selected Services",
```

```
       x = "Service and Status",
       y = "Churn Rate (%)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Relevant Summary Statistics

Based on the preliminary analysis, the following summary statistics are relevant for understanding customer churn:

- Overall churn rate: Approximately 26.5% of customers have churned
- Demographic statistics:
  - Age distribution shows typical consumer age range (18-80 years)
  - Geographic distribution across multiple cities and zip codes
- Service adoption statistics:
  - Internet service types (Fiber Optic, DSL, None)
  - Additional service adoption rates (security, backup, streaming, etc.)
- Financial metrics:
  - Average monthly charges for churned vs. retained customers
  - Total charges and revenue differences between customer groups
- Contract and tenure statistics:
  - Contract type distribution shows higher churn for month-to-month contracts

   - Average tenure for churned customers is significantly lower (approximately 18 months vs. 38 months for non-churned)

Statistical Methods

Primary Analysis Method: Logistic Regression

Logistic regression is appropriate for this analysis because:

- The response variable (churn) is binary (Yes/No)
- We need to quantify the effect of multiple predictors on churn probability
- We want to obtain interpretable odds ratios for business decision-making
- It can handle both categorical and numerical predictors

```r
# Example of logistic regression model (simplified)
churn_model <- glm(
  churned ~ contract + internet_type + tenure_in_months + monthly_charge +
           online_security + premium_tech_support,
  family = binomial(link = "logit"),
  data = telecom_churn
)

# Model summary
summary(churn_model)
```

```
##
## Call:
## glm(formula = churned ~ contract + internet_type + tenure_in_months +
##     monthly_charge + online_security + premium_tech_support,
##     family = binomial(link = "logit"), data = telecom_churn)
##
## Coefficients: (2 not defined because of singularities)
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.629664   0.112604 -14.472  < 2e-16 ***
## contractOne Year        -1.225364   0.102032 -12.010  < 2e-16 ***
## contractTwo Year        -2.381029   0.162003 -14.697  < 2e-16 ***
## internet_typeCable       0.134585   0.191229   0.704    0.482
## internet_typeDSL        -0.240618   0.177542  -1.355    0.175
## internet_typeFiber Optic  0.257543   0.233153   1.105    0.269
## tenure_in_months        -0.025352   0.002017 -12.567  < 2e-16 ***
## monthly_charge           0.015995   0.002390   6.691 2.21e-11 ***
## online_securityNo        0.588100   0.084797   6.935 4.05e-12 ***
## online_securityYes             NA         NA      NA       NA
## premium_tech_supportNo   0.486472   0.085849   5.667 1.46e-08 ***
## premium_tech_supportYes        NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8150.1  on 7042  degrees of freedom
## Residual deviance: 5762.3  on 7033  degrees of freedom
```

```
## AIC: 5782.3
##
## Number of Fisher Scoring iterations: 6
```

```
# Example prediction
predicted_probs <- predict(churn_model, type = "response")
telecom_churn$predicted_churn_prob <- predicted_probs

# ROC curve assessment
roc_obj <- roc(telecom_churn$churned, predicted_probs)
```

```
## Setting levels: control = No, case = Yes
```

```
## Setting direction: controls < cases
```

```
auc_value <- auc(roc_obj)
cat("AUC:", auc_value)
```

```
## AUC: 0.8539964
```

Secondary Analysis Methods

1. **Random Forest Classification**
    - Will help identify complex non-linear relationships and interactions
    - Provides feature importance to highlight the most predictive variables
    - Handles mixed data types effectively

```
# Random Forest example (not evaluated to save computation time)
set.seed(123)
rf_model <- randomForest(
  churned ~ contract + internet_type + tenure_in_months + monthly_charge +
            online_security + premium_tech_support + payment_method + age,
  data = telecom_churn,
  ntree = 100,
  importance = TRUE
)

# Variable importance
varImpPlot(rf_model)
```

2. **Survival Analysis**
    - Can analyze time-to-churn based on tenure
    - Provides insights into when customers are most at risk of churning
    - Allows for censored observations (current customers who haven't churned yet)

Model Evaluation Strategy

The models will be evaluated using:

- Train/test split (70%/30%) for model validation
- Cross-validation to ensure model robustness
- ROC curves and AUC for classification performance
- Confusion matrix for precision, recall, and F1-score
- McFadden's R^2 for logistic regression fit assessment

Expected Outcomes

This analysis is expected to:

1. Identify the key predictors of customer churn in the telecom industry
2. Quantify the impact of each factor on churn probability
3. Develop a predictive model to identify at-risk customers before they churn
4. Provide actionable insights for reducing churn through targeted interventions
5. Generate recommendations for service improvements and retention strategies

The results will be valuable for:

- Marketing teams designing retention campaigns
- Product managers prioritizing service improvements
- Customer service teams implementing proactive retention measures
- Business leaders making strategic decisions about service offerings