MACT6100 Assessment 2

Junchi Han

2025-12-04

1. Introduction

1.1 Introduction to Customer Churn

Customer churn is a major challenge in the telecommunications industry due to intense competition and low switching costs. Existing research shows that churn rates are directly related to the revenue. It is also suggested that retaining customers is more cost-effective than acquiring new ones. As a result, churn prediction has become a key application of machine learning and statistical modelling in both academia and industry.

1.2 Background of Customer Churn Data

Previous academic studies has demonstrated that classification methods can effectively predict customer churn using behavioural and demographic data. This study uses the Telco Customer Churn dataset from Kaggle, which contains rich information on customer demographics, service usage, contracts, and billing behaviour. The topic was chosen due to its strong real-world relevance and its suitability for comparing machine-learning techniques in a practical setting.

1.3 Aim of the Research

This study aims to address the following core questions: how do different models (GLM, Decision Tree, Neural Network, and Random Forest) compare in their performance for churn prediction, and which model performs best? In addition, within the Telco Customer Churn dataset, which features are most important for predicting whether a customer will churn?

1.4 Methology & Models

To address the research questions, this project applies a range of statistical and machine learning methods to build and compare predictive models.

The methodology includes data cleaning, exploratory data analysis (EDA), train-test split, and model performance evaluation.

The models considered include the traditional statistical approach of GLM, as well as machine learning and deep learning models such as Decision Tree, Neural Network, and Random Forest. All models are evaluated and compared using performance metrics including ROC, AUC and accuracy, to provide a comprehensive assessment of their effectiveness in customer churn prediction.

2. Data Loading & Exploring

2.1 Load Customer Churn Data

"Telco Customer Churn" data is downloaded from https://www.kaggle.com/datasets/blastchar/telco-customer-churn?resource=download

```
library(tidyverse)
sdata = read.csv('WA_Fn-UseC_-Telco-Customer-Churn.csv')
str(sdata)
```

```
## 'data.frame':    7043 obs. of  21 variables:
##  $ customerID      : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
##  $ gender          : chr  "Female" "Male" "Male" "Male" ...
##  $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Partner         : chr  "Yes" "No" "No" "No" ...
##  $ Dependents      : chr  "No" "No" "No" "No" ...
##  $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
##  $ PhoneService    : chr  "No" "Yes" "Yes" "No" ...
##  $ MultipleLines   : chr  "No phone service" "No" "No" "No phone service" ...
##  $ InternetService : chr  "DSL" "DSL" "DSL" "DSL" ...
##  $ OnlineSecurity  : chr  "No" "Yes" "Yes" "Yes" ...
##  $ OnlineBackup    : chr  "Yes" "No" "Yes" "No" ...
##  $ DeviceProtection: chr  "No" "Yes" "No" "Yes" ...
##  $ TechSupport     : chr  "No" "No" "No" "Yes" ...
##  $ StreamingTV     : chr  "No" "No" "No" "No" ...
##  $ StreamingMovies : chr  "No" "No" "No" "No" ...
##  $ Contract        : chr  "Month-to-month" "One year" "Month-to-month" "One year" ...
##  $ PaperlessBilling: chr  "Yes" "No" "Yes" "No" ...
##  $ PaymentMethod   : chr  "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic
##  $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
##  $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
##  $ Churn           : chr  "No" "No" "Yes" "No" ...
```

It is suggested that there are 7043 observations of 21 variables in the original churn data. The data are stored in different categories such as character, number or integer.

It is also worthy check if data contains any N/A values.

```
sum(is.na(sdata$TotalCharges))
```

```
## [1] 11
```

```
sdata = na.omit(sdata)
sum(is.na(sdata$TotalCharges))
```

```
## [1] 0
```

Noticed that 11 N/A values have been removed.

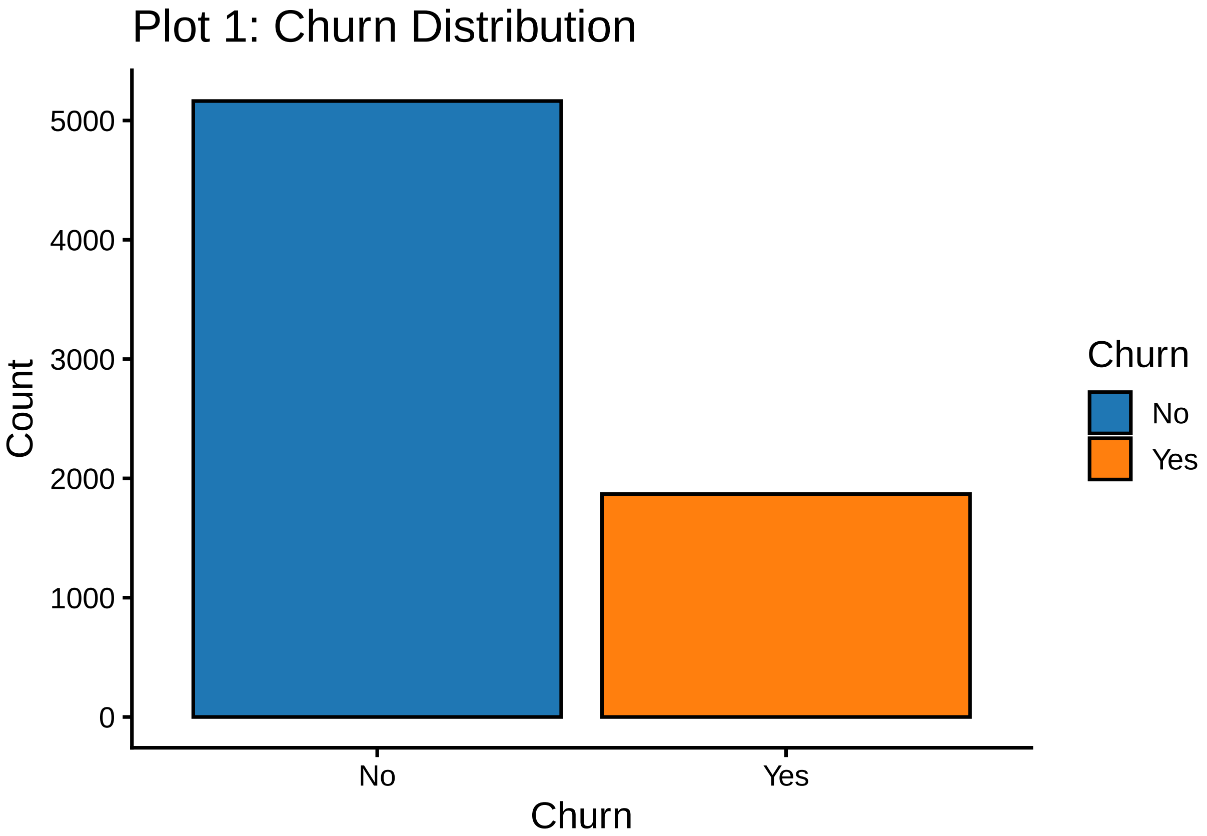
It is sensible to remove "customerID" column, as they are unique identifier of each of the customers and are not useful for prediction.

```
sdata <- sdata %>% dplyr::select(-customerID)
```

2.2 Exploratory Data Analysis on Churn Data

After initial data explore, we want to have understanding of how the elements can be related to each other. Graph is a good method to help to build understanding by visualising the correlations.

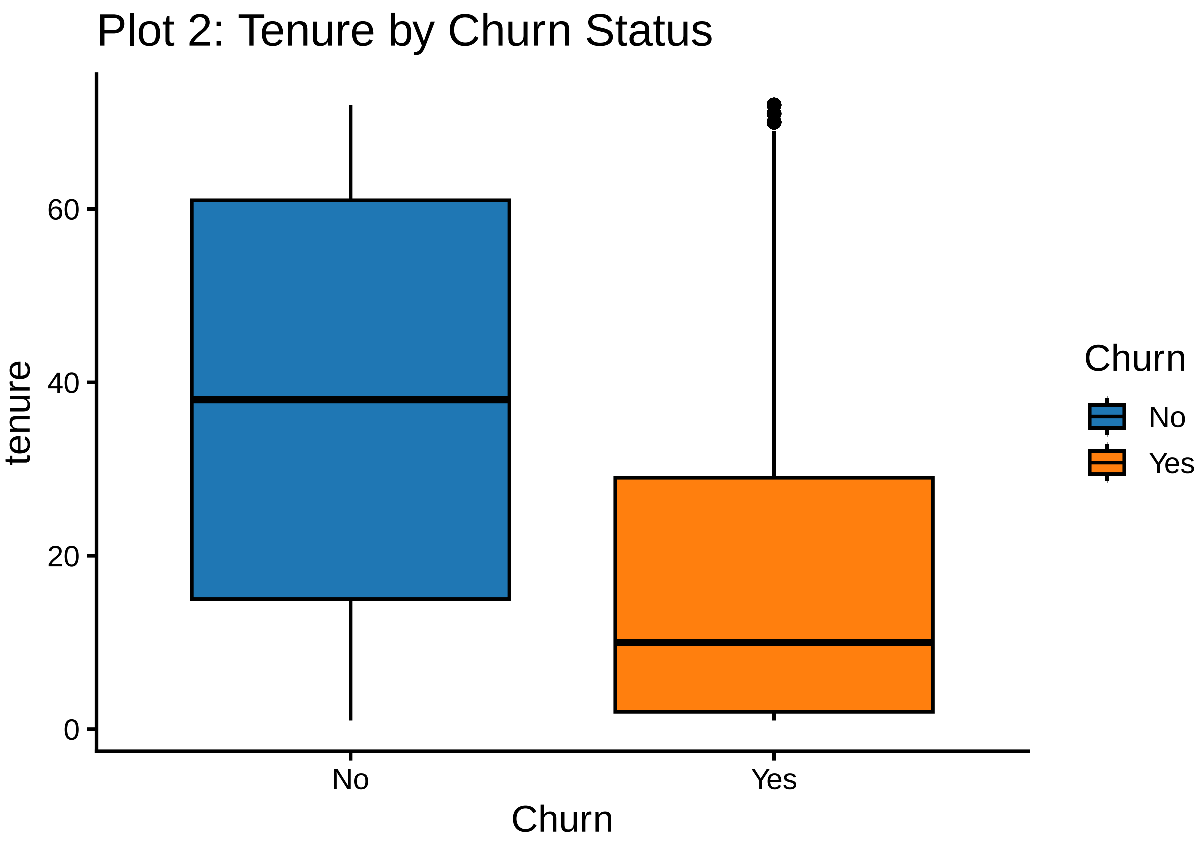
```
library(ggplot2)
library(corrplot)

ggplot(sdata, aes(Churn, fill = Churn)) +
  geom_bar(color="black") +
   scale_fill_manual(values=c("No"="#1f77b4", "Yes"="#ff7f0e")) +
  labs(title = "Plot 1: Churn Distribution",
       y = "Count") +
  theme_classic(base_size = 14)
```



Plot 1:

- The amount of customer classified as “non-churned” is almost triple than those classified as “churn”.

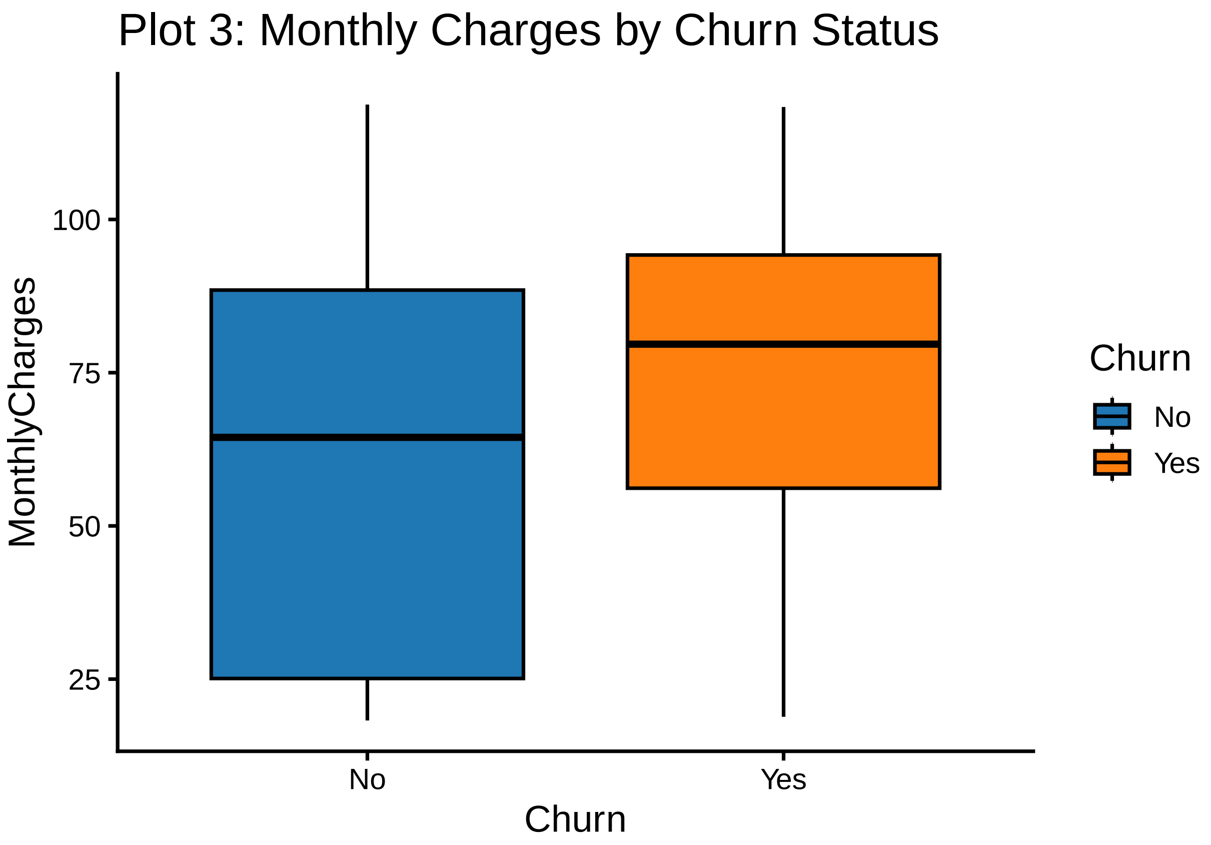
```
ggplot(sdata, aes(Churn, tenure, fill = Churn)) +
  geom_boxplot(color="black") +
  scale_fill_manual(values=c("No"="#1f77b4", "Yes"="#ff7f0e")) +
  labs(title = "Plot 2: Tenure by Churn Status") +
  theme_classic(base_size = 14)
```



Plot 2:

- The tenure distribution of churned customers (Yes) is clearly lower, with most concentrated among short-term users (0–20 months). In contrast, non-churned customers generally have much longer usage periods, and long-term customers have a lower risk of churn.
- This indicates that the shorter the usage period, the more likely customers are to churn, reflecting the characteristics of "new users being less stable and having lower loyalty".

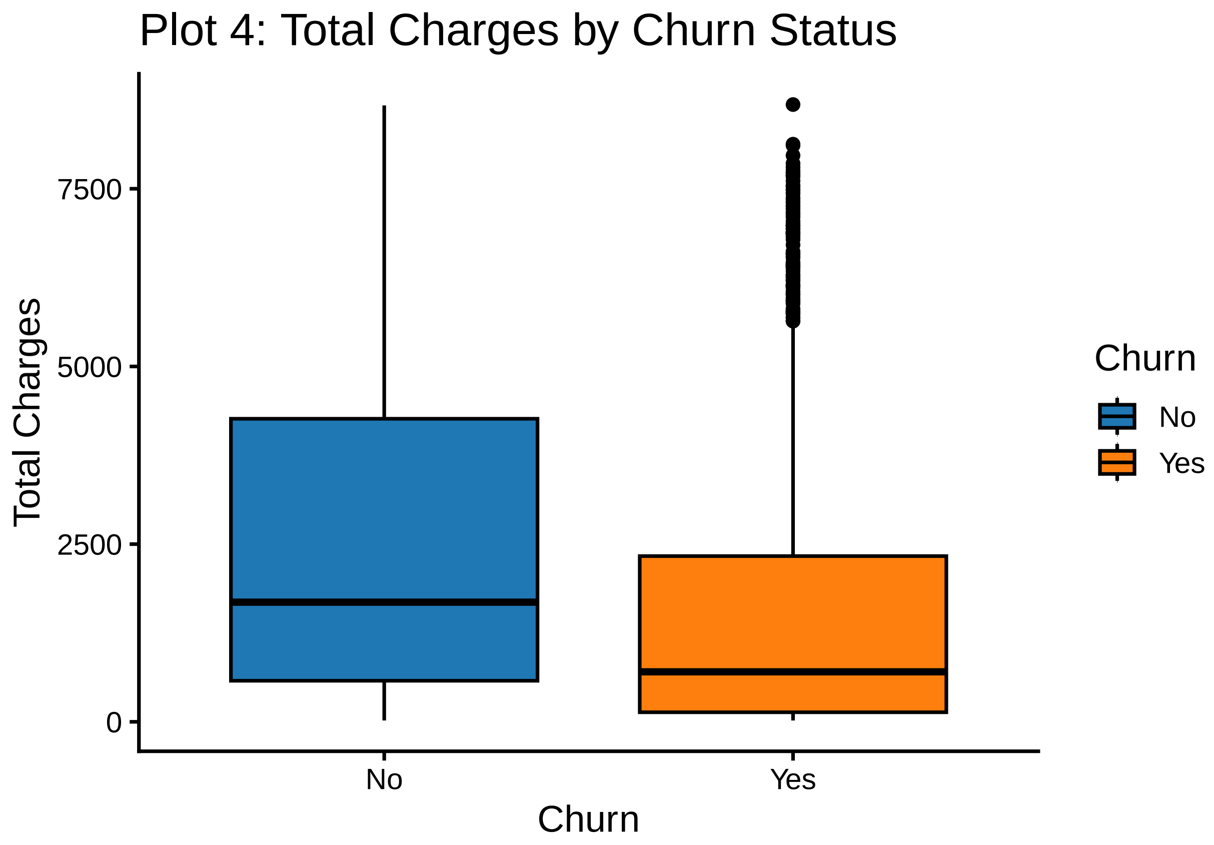
```
ggplot(sdata, aes(Churn, MonthlyCharges, fill = Churn)) +
  geom_boxplot(color="black") +
  scale_fill_manual(values=c("No"="#1f77b4", "Yes"="#ff7f0e")) +
  labs(title = "Plot 3: Monthly Charges by Churn Status") +
  theme_classic(base_size = 14)
```



Plot 3:

- Customers who churn tend to have higher monthly charges than those who remain, as shown by the higher median for the churned group.
- Although there is some overlap between the two groups, the overall distribution is shifted upward for churned customers, suggesting that higher prices are associated with a greater likelihood of churn.

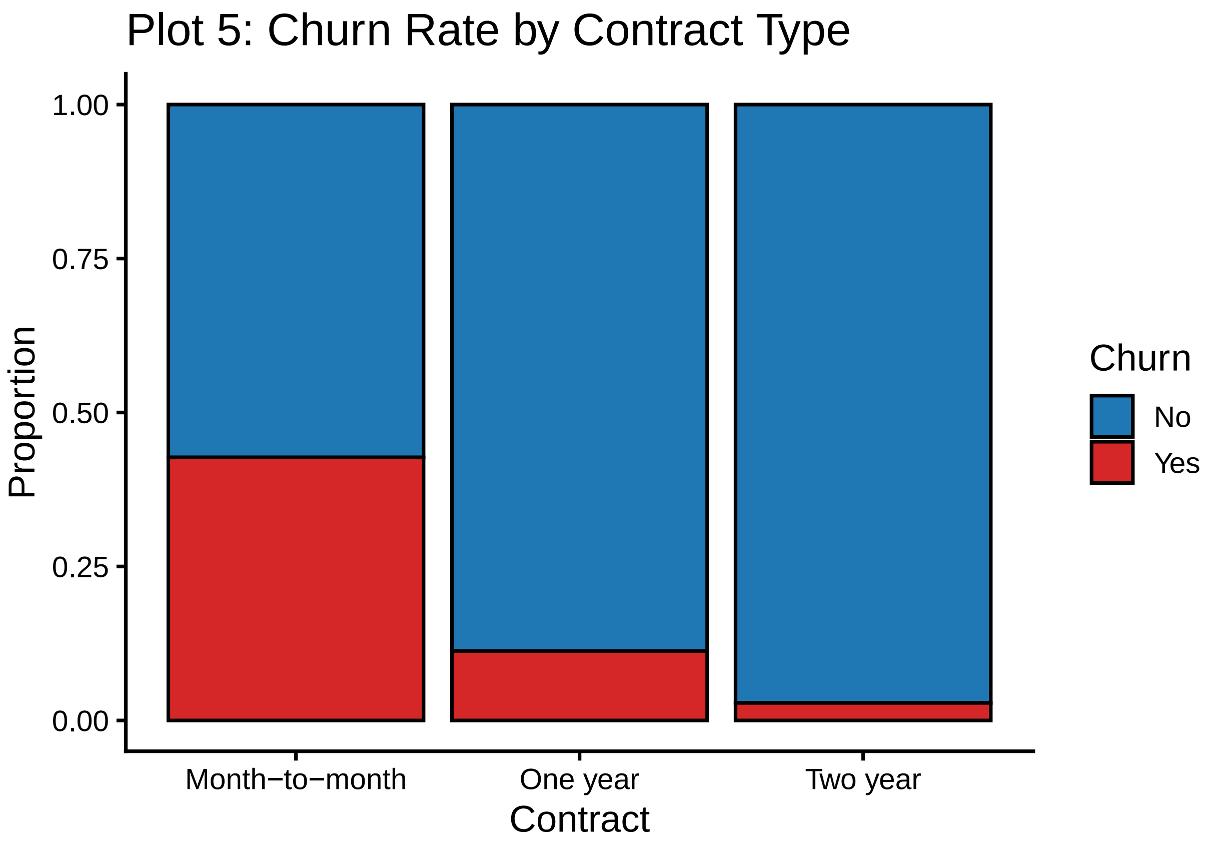
```r
ggplot(sdata, aes(Churn, TotalCharges, fill = Churn)) +
  geom_boxplot(color="black") +
  scale_fill_manual(values=c("No"="#1f77b4", "Yes"="#ff7f0e")) +
  labs(title = "Plot 4: Total Charges by Churn Status",
       x = "Churn",
       y = "Total Charges") +
  theme_classic(base_size = 14)
```



Plot 4:

- Customers who did not churn have much higher total charges on average, reflecting longer tenure, while churned customers generally show low total charges, indicating early-stage departure.
- However, the churned group contains several high-value outliers, showing that although rare, some long-standing, high-spending customers still choose to leave, possibly due to dissatisfaction or external competition.

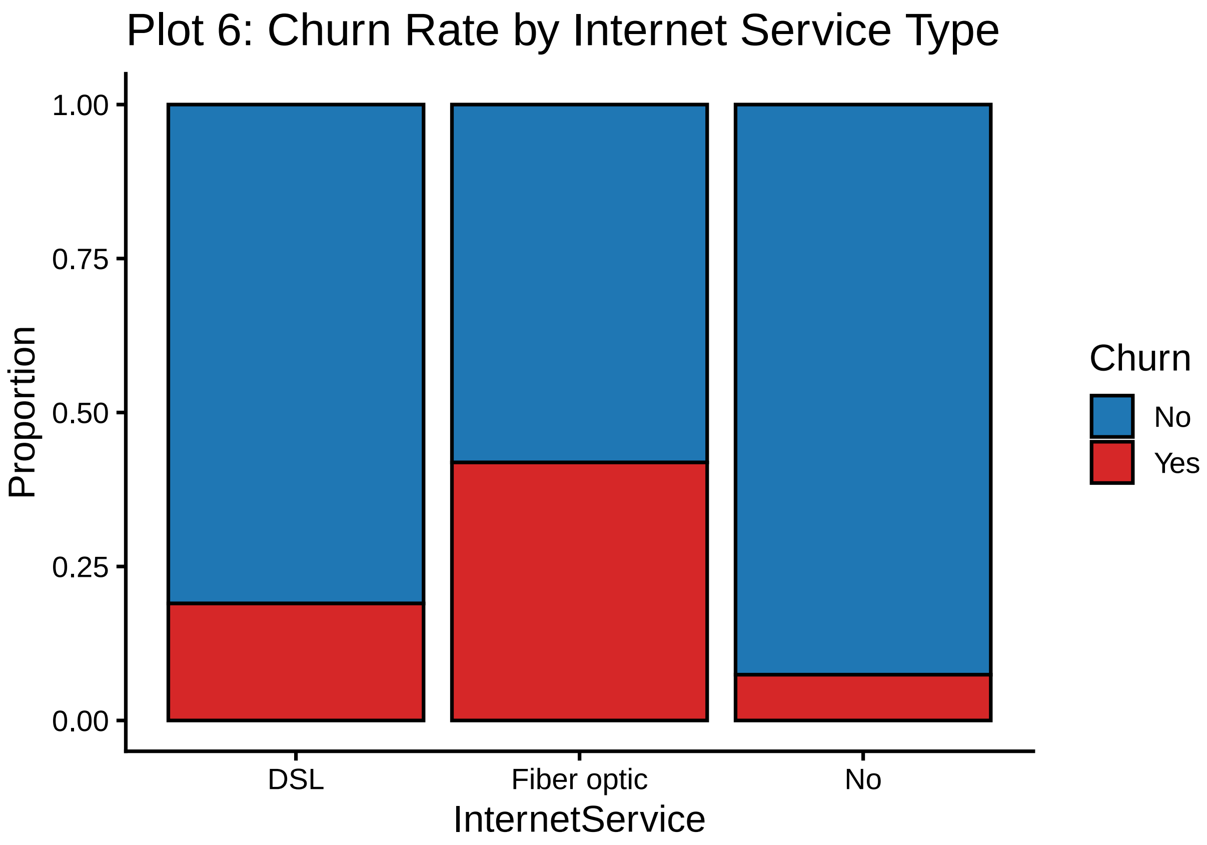
```r
ggplot(sdata, aes(Contract, fill = Churn)) +
  geom_bar(position = "fill", color="black") +
  scale_fill_manual(values=c("No"="#1f77b4", "Yes"="#d62728")) +
  labs(title = "Plot 5: Churn Rate by Contract Type",
       y = "Proportion") +
  theme_classic(base_size = 14)
```



Plot 5:

- Customers on month-to-month contracts exhibit the highest churn rate, while those on one-year and especially two-year contracts are far more likely to remain.
- This indicates that longer contract durations are strongly associated with improved customer retention, highlighting contract length as a key driver of churn.

```
ggplot(sdata, aes(InternetService, fill = Churn)) +
  geom_bar(position = "fill", color="black") +
  scale_fill_manual(values=c("No"="#1f77b4", "Yes"="#d62728")) +
  labs(title = "Plot 6: Churn Rate by Internet Service Type",
       y = "Proportion") +
  theme_classic(base_size = 14)
```


Plot 6: Churn Rate by Internet Service Type
Proportion
1.00
0.75
0.50
0.25
0.00
DSL
Fiber optic
No
InternetService
Churn
🟦 No
🟥 Yes