CHAPTER - I

INTRODUCTION

TELECOM CHURN ANALYSIS

1.1 Scope of analysis

- The primary objective of this analysis is to identify the factors contributing to customer churn and build a predictive model to classify customers as churners or non-churners. The insights derived will help the telecom company implement targeted retention strategies to minimize churn and improve customer satisfaction.
- The data was downloaded from IBM Sample Data Sets for customer retention programs. The goal of this project is to predict behaviours of churn or not churn to help retain customers. Each row represents a customer, each column contains a customer's attribute.
- Customers who left within the last month – the column is called Churn Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges Demographic info about customers – gender, age range, and if they have partners and dependents

1.2 Approach of Analysis

The approach to analysis the Telco Customer Churn dataset involves cleaning the data to handle missing values and inconsistencies, followed by exploratory data analysis (EDA) to identify trends and correlations between customer churn and features like tenure, contract type, and monthly charges. Relevant features are transformed and encoded for machine learning models, addressing class imbalance through techniques like oversampling. Predictive models such as Logistic Regression, Random Forest, SVM and KNN are built and evaluated using metrics like accuracy and recall. Insights from the analysis guide actionable recommendations to reduce churn and improve customer retention strategies.

CHAPTER II

DATA UNDERSTANDING

TELECOM CHURN ANALYSIS

2.1 Data Understanding

Load the relevant Packages

```{r}
suppressMessages(library(tidyverse))
suppressMessages(library(caret))
suppressMessages(library(reshape2))
suppressMessages(library(broom))
suppressMessages(library(randomForest))
suppressMessages(library(performanceEstimation))
suppressMessages(library(regclass))
suppressMessages(library(GGally))
suppressMessages(library(pROC))
suppressMessages(library(plotROC))
suppressMessages(library(cowplot))
suppressMessages(library(grid))
suppressMessages(library(gridExtra))
suppressMessages(library(formattable))
suppressMessages(library(scales))
suppressMessages(library(ggplot2))
library(kernlab)
theme_set(theme_minimal())
options(warn=-1)
```

Load the dataset

```{r}

telecom <- read_csv("C:/Users/Deepak/Documents/Project details/WA_Fn-UseC_-Telco-Customer-Churn.csv")

view(telecom)

```

Structure of the data

```{r}

str(telecom)

```

```
spc_tbl_ [7,043 × 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
(automatic)" ...
 $ 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" ...
```

2.2 Data description

The telecom dataset has 7043 rows and 21 columns

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	Onli
1	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No
2	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes
3	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes
4	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes
5	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No
6	9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No
7	1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No
8	6713-OKOMC	Female	0	No	No	10	No	No phone service	DSL	Yes
9	7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No
10	6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes
11	9763-GRSKD	Male	0	Yes	Yes	13	Yes	No	DSL	Yes
12	7469-LKBCI	Male	0	No	No	16	Yes	No	No	No i
13	8091-TTVAX	Male	0	Yes	No	58	Yes	Yes	Fiber optic	No
14	0280-XJGEX	Male	0	No	No	49	Yes	Yes	Fiber optic	No
15	5129-JLPIS	Male	0	No	No	25	Yes	No	Fiber optic	Yes
16	3655-SNQYZ	Female	0	Yes	Yes	69	Yes	Yes	Fiber optic	Yes
17	8191-XWSZG	Female	0	No	No	52	Yes	No	No	No i
18	9959-WOFKT	Male	0	No	Yes	71	Yes	Yes	Fiber optic	Yes
19	4190-MFLUW	Female	0	Yes	Yes	10	Yes	No	DSL	No
20	4183-MYFRB	Female	0	No	No	21	Yes	No	Fiber optic	No
21	8779-QRDMV	Male	1	No	No	1	No	No phone service	DSL	No
22	1680-VDCWW	Male	0	Yes	No	12	Yes	No	No	No i
23	1066-JKSGK	Male	0	No	No	1	Yes	No	No	No i
24	3638-WEABW	Female	0	Yes	No	58	Yes	Yes	DSL	No
25	6322-HRPFA	Male	0	Yes	Yes	49	Yes	No	DSL	Yes
26	6865-JZNKO	Female	0	No	No	30	Yes	No	DSL	Yes
27	6467-CHFZW	Male	0	Yes	Yes	47	Yes	Yes	Fiber optic	No

Here is an explanation of the common variables in the Telco Customer Churn dataset:

1. Demographic Variables

These describe the customer's personal details.

Gender

The gender of the customer (e.g., Male, Female).

Senior Citizen

Indicates if the customer is a senior citizen (0 = No, 1 = Yes).

Partner

Indicates if the customer has a partner (Yes or No).

Dependents

Indicates if the customer has dependents (Yes or No).

2. Services Variables

These describe the telecom services subscribed to by the customer.

Phone Service

Indicates if the customer has a phone service (Yes or No).

Multiple Lines

Indicates if the customer has multiple phone lines (No, Yes, or No phone service).

Internet Service

The type of internet service (e.g., DSL, Fibre optic, or No).

Online Security

Indicates if the customer has online security add-ons (Yes, No, or No internet service).

Online Backup

Indicates if the customer has online backup add-ons (Yes, No, or No internet service).

Device Protection

Indicates if the customer has device protection add-ons (Yes, No, or No internet service).

Tech Support

Indicates if the customer has technical support add-ons (Yes, No, or No internet service).

Streaming TV

Indicates if the customer uses streaming TV services (Yes, No, or No internet service).

Streaming Movies

Indicates if the customer uses streaming movies services (Yes, No, or No internet service).

3. Account Variables

These describe the customer's account details.

Contract

The type of contract (e.g., Month-to-month, One year, Two year).

Paperless Billing

Indicates if the customer has opted for paperless billing (Yes or No).

Payment Method

The customer's payment method (e.g., Electronic check, Mailed check, Bank transfer, Credit card).

Monthly Charges

The monthly charge for the customer (numeric).

Total Charges

The total amount billed to the customer (numeric, sometimes has missing or inconsistent values).

Tenure:

The number of months the customer has been with the company (numeric).

4. Target Variable

This variable indicates whether the customer has churned.

Churn

The target variable, showing whether the customer has left the company (Yes or No).

This module explains data understanding. This dataset consist of different columns. Each and every columns we should find the summary () function. This function is used to calculate the average value and determine the maximum, minimum of the column in a data frame.

````
```{r}
summary(telecom)
```
````

```
   customerID           gender          SeniorCitizen      Partner           Dependents
 Length:7043        Length:7043        Min.   :0.0000   Length:7043        Length:7043
 Class :character   Class :character   1st Qu.:0.0000   Class :character   Class :character
 Mode  :character   Mode  :character   Median :0.0000   Mode  :character   Mode  :character
                                       Mean   :0.1621
                                       3rd Qu.:0.0000
                                       Max.   :1.0000

     tenure     PhoneService       MultipleLines      InternetService    OnlineSecurity
 Min.   : 0.00   Length:7043        Length:7043        Length:7043        Length:7043
 1st Qu.: 9.00   Class :character   Class :character   Class :character   Class :character
 Median :29.00   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :32.37
 3rd Qu.:55.00
 Max.   :72.00

 OnlineBackup       DeviceProtection   TechSupport        StreamingTV        StreamingMovies
 Length:7043        Length:7043        Length:7043        Length:7043        Length:7043
 Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character



   Contract         PaperlessBilling   PaymentMethod      MonthlyCharges    TotalCharges
 Length:7043        Length:7043        Length:7043        Min.   : 18.25   Min.   :  18.8
 Class :character   Class :character   Class :character   1st Qu.: 35.50   1st Qu.: 401.4
 Mode  :character   Mode  :character   Mode  :character   Median : 70.35   Median :1397.5
                                                          Mean   : 64.76   Mean   :2283.3
                                                          3rd Qu.: 89.85   3rd Qu.:3794.7
                                                          Max.   :118.75   Max.   :8684.8
                                                                           NA's   :11
    Churn
 Length:7043
 Class :character
 Mode  :character
```

2.3 Handle Missing Values

Check for missing values

```{r}

# Check for missing values

colSums(is.na(telecom))

```

```
      customerID           gender    SeniorCitizen          Partner       Dependents           tenure
               0                0                0                0                0                0
    PhoneService    MultipleLines  InternetService   OnlineSecurity     OnlineBackup DeviceProtection
               0                0                0                0                0                0
     TechSupport      StreamingTV  StreamingMovies         Contract PaperlessBilling    PaymentMethod
               0                0                0                0                0                0
  MonthlyCharges     TotalCharges            Churn
               0               11                0
```

Convert to numeric and remove na values in dataset

```{r}
#convert to numeric and remove na values in dataset
telecom$TotalCharges <- as.numeric(as.character(telecom$TotalCharges))
telecom <- telecom %>% na.omit()
```

Again check is there NA value in data set

```{r}
#again check is there NA value in data set
colSums(is.na(telecom))
```

```
      customerID           gender    SeniorCitizen          Partner       Dependents           tenure 
               0                0                0                0                0                0 
    PhoneService    MultipleLines  InternetService   OnlineSecurity     OnlineBackup DeviceProtection 
               0                0                0                0                0                0 
     TechSupport      StreamingTV  StreamingMovies         Contract PaperlessBilling    PaymentMethod 
               0                0                0                0                0                0 
  MonthlyCharges     TotalCharges            Churn 
               0                0                0 
```

2.4 Explore categorical variables

```{r}
# Identify categorical variables
categorical_vars <- telecom %>%
  select_if(is.character)
```

Remove the unique variable from categorical variable

```{r}
# Remove the customerID column from categorical variables for shot the output it is also a categorical variable
categorical_vars <- categorical_vars %>%
  select(-customerID)
```

Add Senior Citizen column as a categorical variable

```{r}
# Add SeniorCitizen column as a categorical variable
categorical_vars <- telecom %>%
  mutate(SeniorCitizen = as.factor(SeniorCitizen)) %>% # Convert SeniorCitizen to a factor
  select(c(colnames(categorical_vars), "SeniorCitizen")) # Include SeniorCitizen with other categorical variables
```

Display unique categories for each categorical variable

```{r}
# Display unique categories for each categorical variable
lapply(categorical_vars, function(column) {
  unique(column)
})
```

```
$gender
[1] "Female" "Male"

$Partner
[1] "Yes" "No"

$Dependents
[1] "No"  "Yes"

$PhoneService
[1] "No"  "Yes"

$MultipleLines
[1] "No phone service" "No"               "Yes"

$InternetService
[1] "DSL"         "Fiber optic" "No"

$OnlineSecurity
[1] "No"                  "Yes"                 "No internet service"

$OnlineBackup
[1] "Yes"                 "No"                  "No internet service"

$DeviceProtection
[1] "No"                  "Yes"                 "No internet service"

$TechSupport
[1] "No"                  "Yes"                 "No internet service"

$StreamingTV
[1] "No"                  "Yes"                 "No internet service"

$StreamingMovies
[1] "No"                  "Yes"                 "No internet service"

$Contract
[1] "Month-to-month" "One year"       "Two year"

$PaperlessBilling
[1] "Yes" "No"

$PaymentMethod
[1] "Electronic check"          "Mailed check"              "Bank transfer (automatic)"
[4] "Credit card (automatic)"

$Churn
[1] "No"  "Yes"

$SeniorCitizen
[1] 0 1
Levels: 0 1
```

Remove unwanted column customer id and Convert character variables to factors

```{r}
# Remove unnessary column customer id and Convert character variables to factors
telecom <- telecom %>%
  select(-customerID)%>%
  mutate_at(7, ~as.factor(case_when(. == "No phone service" ~ "No",
                                    . == "No" ~ "No", TRUE ~ "Yes"))) %>%
  mutate_at(8, ~as.factor(case_when(. == "Fibre optic" ~ "FibreOptic",
                                    . == "DSL" ~ "DSL", TRUE ~ "No"))) %>%
  mutate_at(c(9:14), ~as.factor(case_when(. == "No internet service" ~ "No",
                                          . == "No" ~ "No", TRUE ~ "Yes"))) %>%
  mutate_at(17, ~as.factor(case_when(. == "Bank transfer (automatic)" ~
"BankTransferAuto",
                                     . == "Credit card (automatic)" ~ "CreditCardAuto",
                                     . == "Electronic check" ~ "ECheck", TRUE ~
"MailedCheck")))
```

Convert character variables to factors

```{r}
# Convert character variables to factors
telecom <- telecom %>%
  mutate(across(where(is.character), as.factor))
```

Summary statistics by gender

```{r}
telecom %>%
  group_by(gender) %>%
  rename("Gender" = gender) %>%
  summarise("Number of Observations" = n(),
            "Average Tenure, in months" = round(mean(tenure), 0),
            "Monthly Charges" = round(mean(MonthlyCharges), 2))
```

	Gender	Number of Observations	Average Tenure, in months	Monthly Charges
1	Female	3483	32	65.22
2	Male	3549	33	64.39

CHAPTER III

PREPARING AND EXPLORING DATA

TELECOM CHURN ANALYSIS

3.1 Data Exploration

- When you first get your data, it's very tempting to immediately begin fitting models and assessing how they perform. However, before you begin modelling, it's absolutely essential to explore the structure of the data and the relationships between the variables in the data set.
- Do a detailed EDA of the data set, to learn about the structure of the data and the relationships between the variables in the data set (refer to Data description sheet of data). Your EDA should involve creating and reviewing many plots/graphs and considering the patterns and relationships you see.

Customer's average tenure with Telco and their average charges

```
```{r}
t2 <- telecom %>%
  mutate(Churn2 = as.factor(ifelse(Churn == "Yes", "Former Customers", "Current
Customers")))

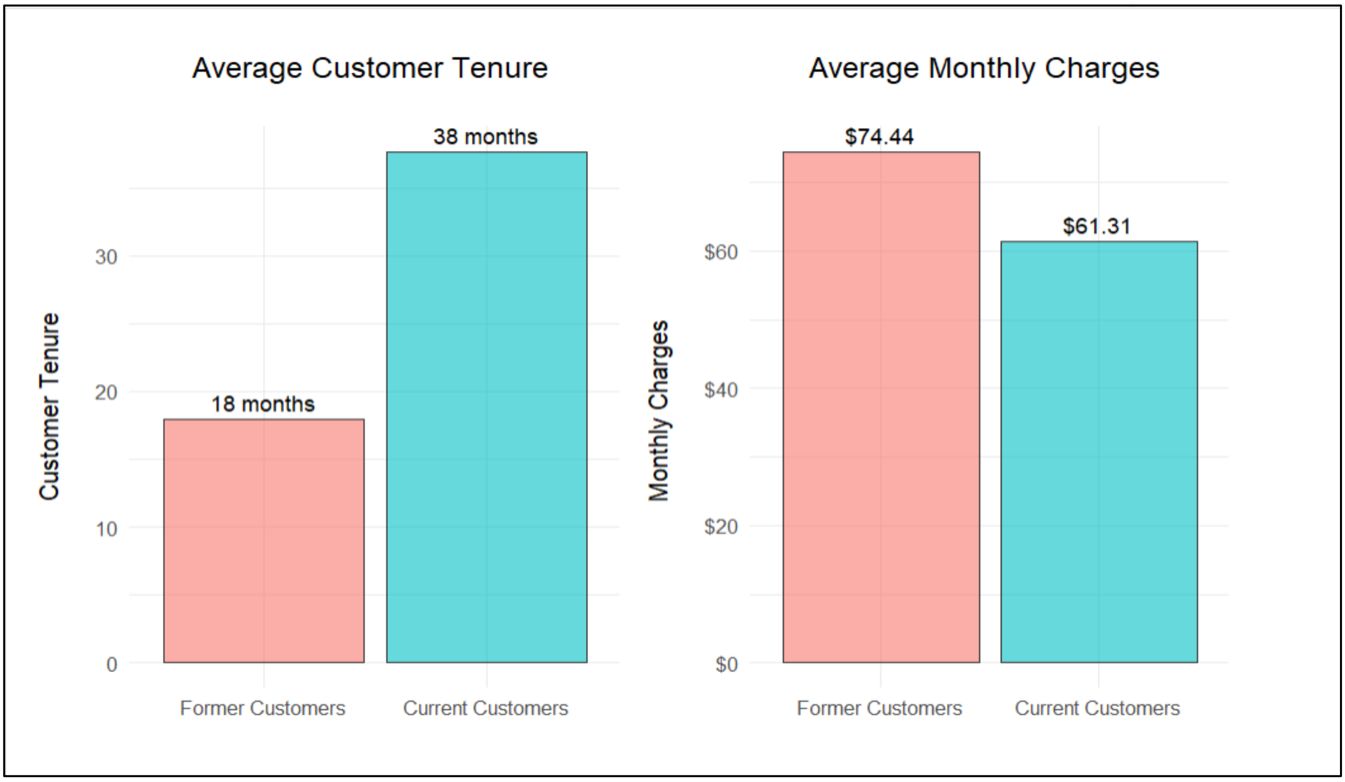

g1 <- ggplot(t2, aes(x = fct_rev(Churn2), y = tenure, fill = fct_rev(Churn2))) +
  geom_bar(stat = "summary", fun = "mean", alpha = 0.6, color = "grey20",
show.legend = F) +
  stat_summary(aes(label = paste(round(..y.., 0), "months")), fun = mean,
             geom = "text", size = 3.5, vjust = -0.5) +
  labs(title = "Average Customer Tenure \n", x = "", y = "Customer Tenure\n") +
  theme(plot.title = element_text(hjust = 0.5))


g2 <- ggplot(t2, aes(x = fct_rev(Churn2), y = MonthlyCharges, fill = fct_rev(Churn2)))
+
  geom_bar(stat = "summary", fun = "mean", alpha = 0.6, color = "grey20",
show.legend = F) +
```
```

```
  stat_summary(aes(label = dollar(..y..)), fun = mean,
               geom = "text", size = 3.5, vjust = -0.5) +
  scale_y_continuous(labels = dollar_format()) +
  labs(title = "Average Monthly Charges \n", x = "", y = "Monthly Charges \n") +
  theme(plot.title = element_text(hjust = 0.5))

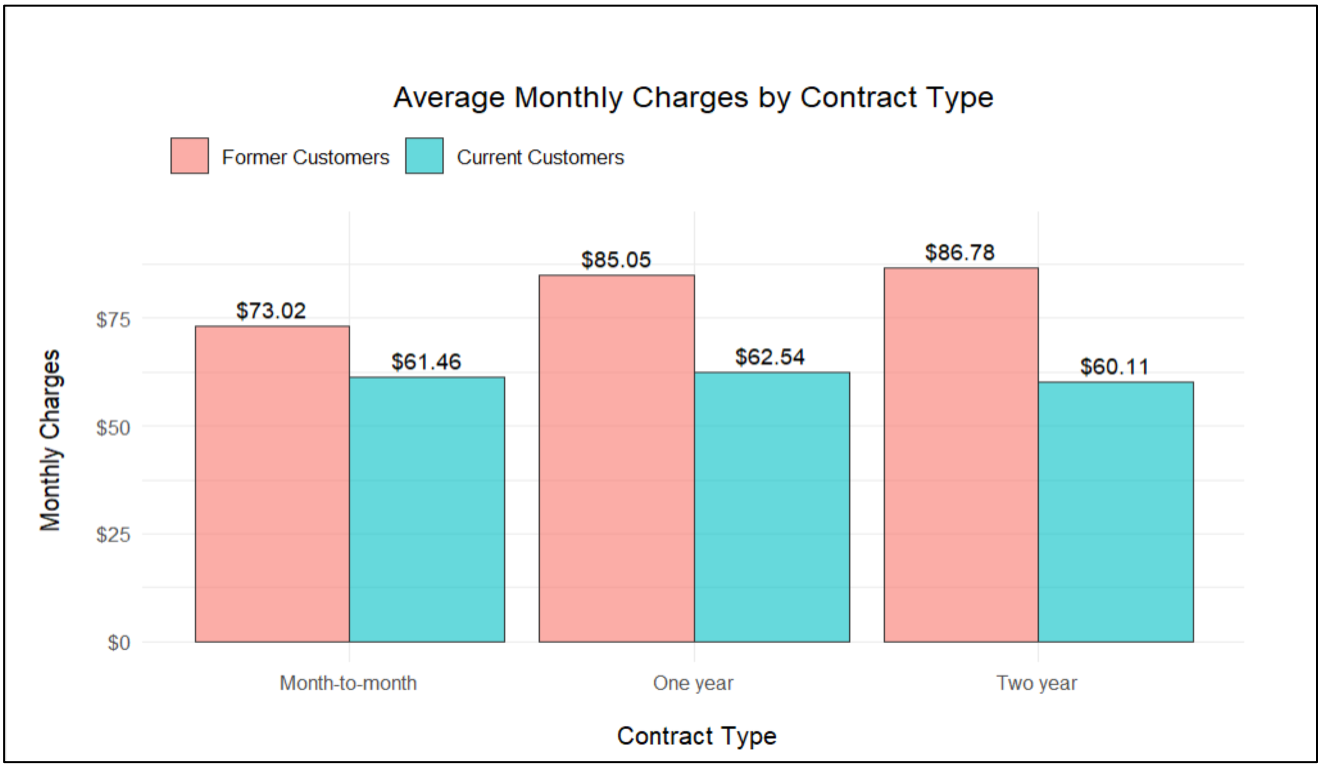
g3 <- ggplot(t2, aes(x = Contract, y = MonthlyCharges, fill = fct_rev(Churn2))) +
  geom_bar(position = "dodge", stat = "summary", fun = "mean", alpha = 0.6, color = "grey20") +
  stat_summary(aes(label = dollar(..y..)), fun = mean,
               geom = "text", size = 3.5, vjust = -0.5,
               position = position_dodge(width = 0.9)) +
  coord_cartesian(ylim = c(0, 95)) +
  scale_y_continuous(labels = dollar_format()) +
  labs(title = "\nAverage Monthly Charges by Contract Type", x = "\n Contract Type",
       y = "Monthly Charges \n", fill = "") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "top", legend.justification = "left")

options(repr.plot.width=10, repr.plot.height=14)
grid.arrange(g1, g2, ncol = 2, nrow = 1, layout_matrix = rbind(c(1,2)))
```



Average Monthly Charges by Contract Type

```{r}
grid.arrange( g3, ncol = 2, nrow = 1, layout_matrix = rbind(c(3,3)))
```



The graphs above show the average tenure of Telco's current and former customers and their monthly charges. Telco's current customers have been with the company for just over 3 years, while customers who left kept their services for about 18 months. Additionally, former customers had higher monthly charges on average by about $13. This holds true across each contract type.

What type of account services do customers having

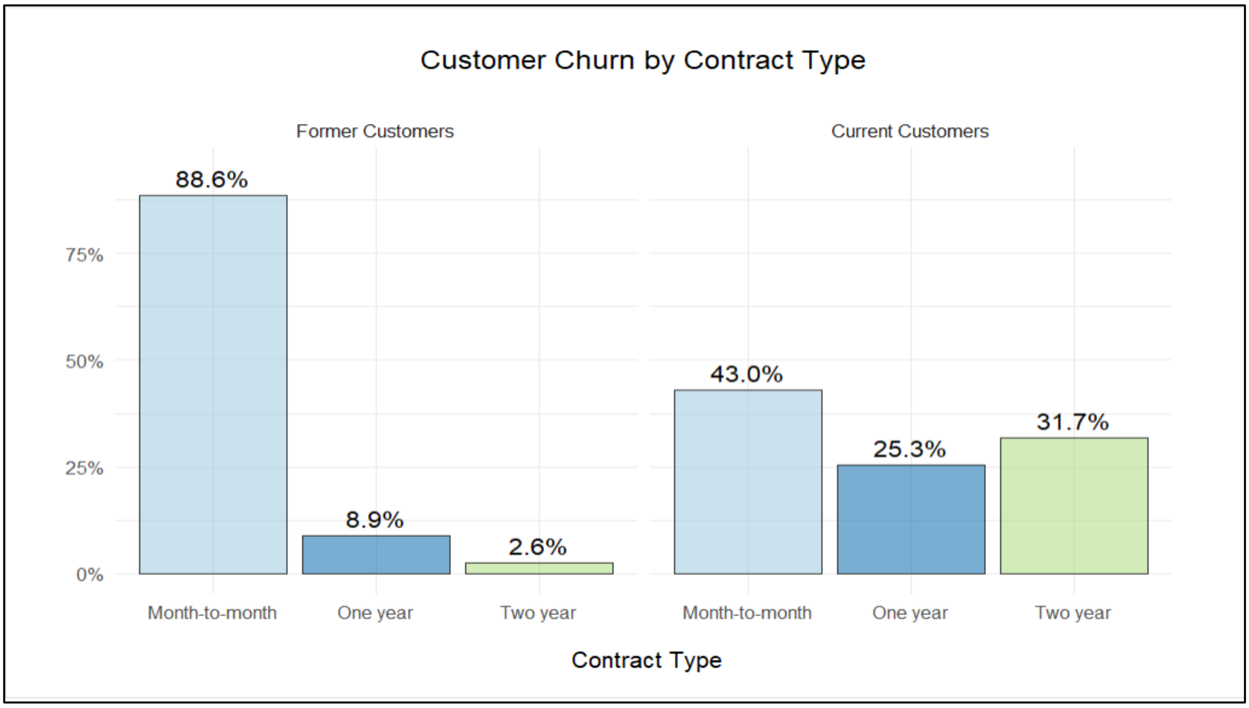
Customer churn by contract type

```{r}
g1 <- ggplot(t2, aes(x = Contract, group = fct_rev(Churn2))) +
  geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat = "count",
          alpha = 0.6, color = "grey20", show.legend = F) +
  geom_text(aes(label = percent(..prop..), y = ..prop.. ),
           size = 4, stat = "count", vjust = -0.5) +
  facet_grid(~fct_rev(Churn2)) +
  scale_y_continuous(labels = percent_format()) +
  coord_cartesian(ylim = c(0, .95)) +
  scale_fill_brewer(palette = "Paired") +
  labs(title = "Customer Churn by Contract Type\n", x = "\n Contract Type", y = "") +
  theme(plot.title = element_text(hjust = 0.5))

g2 <- ggplot(t2, aes(x = InternetService, group = fct_rev(Churn2))) +
  geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat = "count",
          alpha = 0.6, color = "grey20", show.legend = F) +
  geom_text(aes(label = percent(..prop..), y = ..prop.. ),
           size = 4, stat = "count", vjust = -0.5) +
  facet_grid(~fct_rev(Churn2)) +
  scale_y_continuous(labels = percent_format()) +
  coord_cartesian(ylim = c(0, .9)) +
  scale_fill_brewer(palette = "Paired") +
  labs(title = "\n Customer Churn by Internet Service \n", x = "\n Internet Service", y =
"") +
```

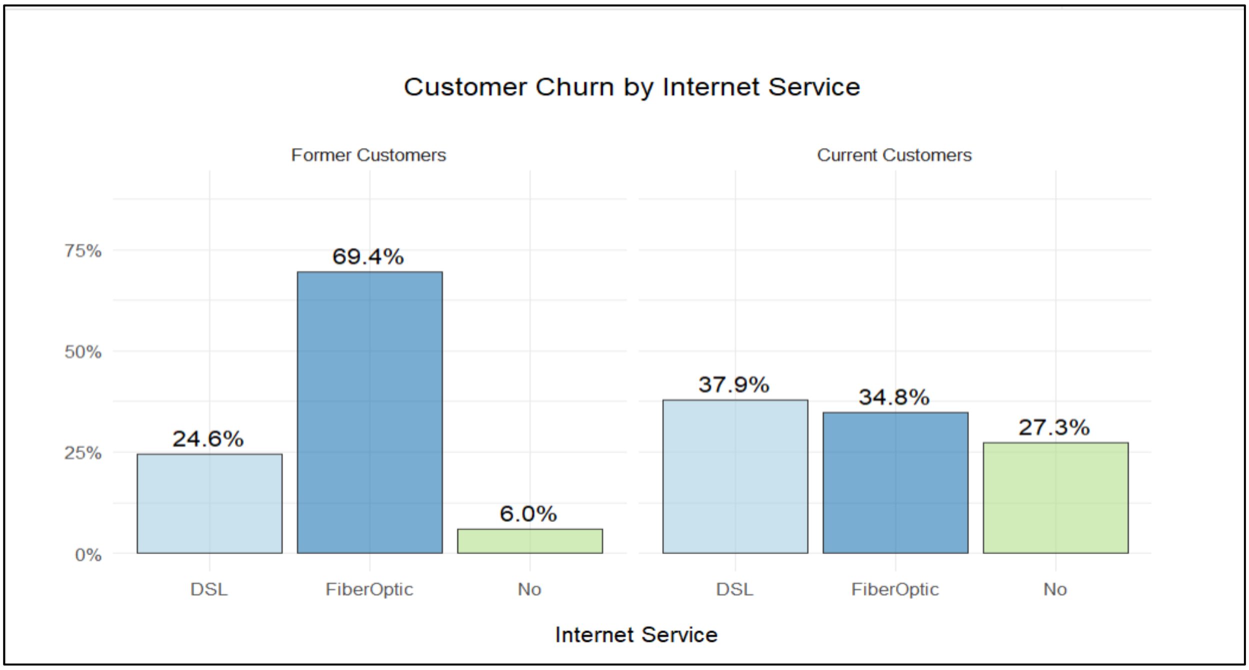
```
theme(plot.title = element_text(hjust = 0.5))

grid.arrange(g1, ncol = 1)
```



Customer churn by internet service

```{r}
grid.arrange(g2, ncol = 1)
```



Nearly 89% of former customers were on month-to-month contracts, with a much smaller proportion in one or two-year contracts. Of customers who left, a little over 69% had Fibre Optic internet. This could be an indicator of potential dissatisfaction with the service and should be further reviewed by the company since currently over a third of their customers have this type of internet.

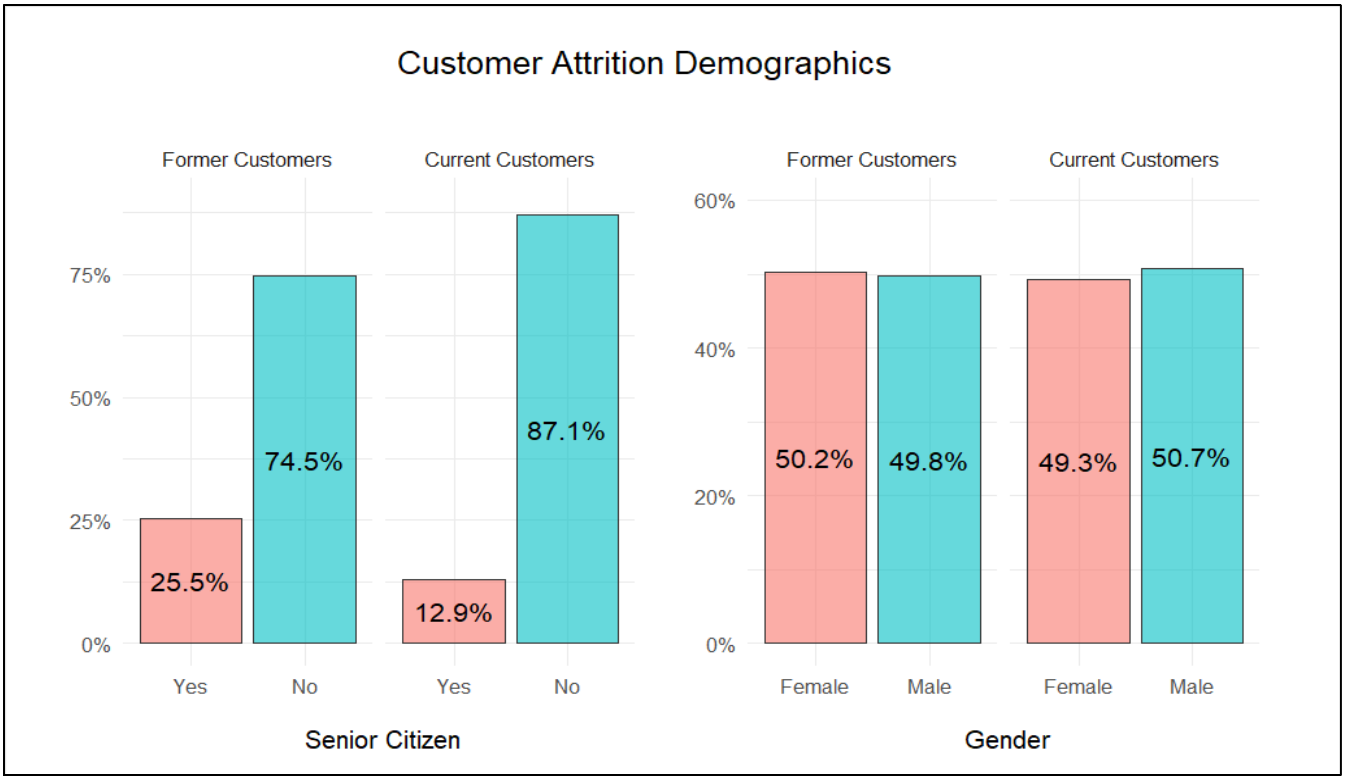
Customer Attrition Demographics

```{r}
g1 <- ggplot(t2, aes(x = fct_rev(ifelse(SeniorCitizen==1, "Yes", "No")), group = Churn2)) +
  geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat = "count",
           alpha = 0.6, color = "grey20", show.legend = F) +
  geom_text(aes(label = percent(..prop.., accuracy = 0.1), y = ..prop..),
            size = 4, stat = "count", position = position_stack(vjust = 0.5)) +
  facet_grid(~fct_rev(Churn2)) +
  scale_y_continuous(labels = percent_format(accuracy = 1)) +
  coord_cartesian(ylim = c(0, .9)) +
  labs(x = "\n Senior Citizen", y = "")

g2 <- ggplot(t2, aes(x = gender, group = Churn2)) +
  geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat = "count",
           alpha = 0.6, color = "grey20", show.legend = F) +
  geom_text(aes(label = percent(..prop.., accuracy = 0.1), y = ..prop..),
            size = 4, stat = "count", position = position_stack(vjust = 0.5)) +
  facet_grid(~fct_rev(Churn2)) +
  scale_y_continuous(labels = percent_format(accuracy = 1)) +
  coord_cartesian(ylim = c(0, .6)) +
  labs(x = "\n Gender", y = "")

options(repr.plot.width=18, repr.plot.height=7)
grid.arrange(g1, g2, nrow = 1, top = textGrob("Customer Attrition Demographics \n",
                                              gp = gpar(fontsize = 14)))
```

````
```
````

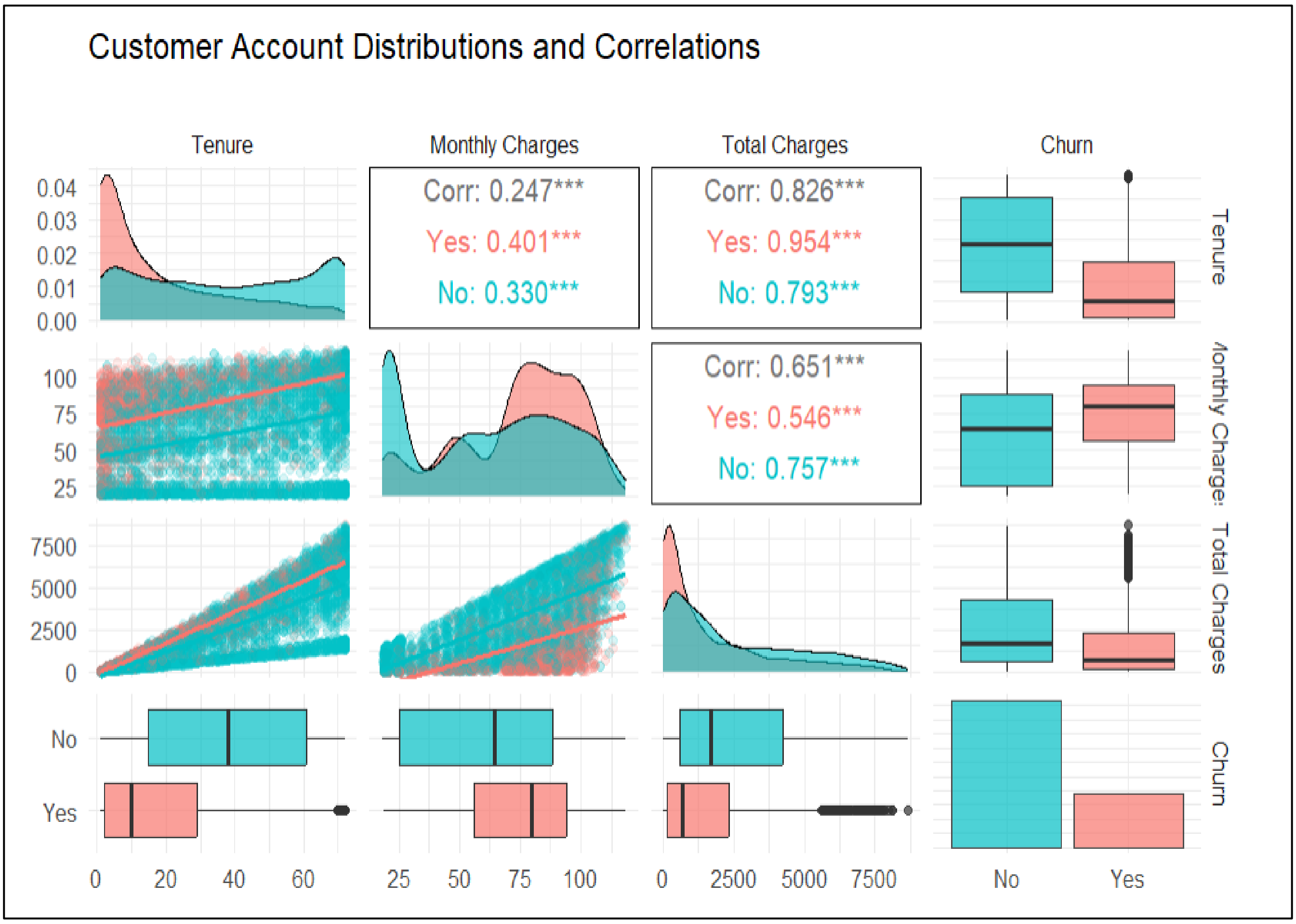


Based on the demographic attributes of Telco's customers, about a quarter of those who left were senior citizens, and just under 13% of their current customers are 65 years or older. The distribution of gender is proportional in both current and former customers, with an approximately equal number of men and women leaving within the last month.

Distributions and Correlations

````
```{r}
options(repr.plot.width=12, repr.plot.height=10)
telecom %>%
  select(tenure, MonthlyCharges, TotalCharges, Churn) %>%
  ggpairs(aes(color = fct_rev(Churn)), title = "Customer Account Distributions and
 Correlations \n",
          columnLabels = c("Tenure", "Monthly Charges", "Total Charges", "Churn"),
          upper = list(combo = wrap("box_no_facet", alpha = 0.7)),
          diag = list(continuous = wrap("densityDiag", alpha = 0.6),
                      discrete = wrap("barDiag", alpha = 0.7, color = "grey30")),
````

```
lower = list(combo = wrap("box_no_facet", alpha = 0.7), continuous =
wrap("smooth", alpha = 0.15)))
```



The correlations between our numeric variables show that TotalCharges is strongly correlated with customer tenure, especially among customers who left (Churn = Yes), with a correlation of more than 0.95. There is also a slightly positive relationship between MonthlyCharges and Tenure of 0.25 and it is significant. The histogram of MonthlyCharges has a unique shape that appears to be multimodal, while the distribution of customer tenure is relatively uniform among current customers but skewed to the right in customers who left.

3.2 Issues in the Dataset

The dataset exhibits a **churn imbalance**, where the number of customers who have churned is significantly lower than those who have not. This class imbalance can negatively impact machine learning models, causing them to favour the majority class (non-churners) and misclassify actual churners. Without addressing this imbalance, the model may have high overall accuracy but poor recall for churn prediction.

3.3 Resolving Issues

To handle the class imbalance, **SMOTE (Synthetic Minority Over-sampling Technique)** is applied. SMOTE generates synthetic samples for the minority class (churners) using nearest-neighbour techniques. This method helps create a balanced dataset, ensuring that models can effectively learn patterns for both churned and non-churned customers. After applying SMOTE, the dataset has an equal distribution of churners and non-churners, improving prediction accuracy and recall.

SMOTE for Churn Balance

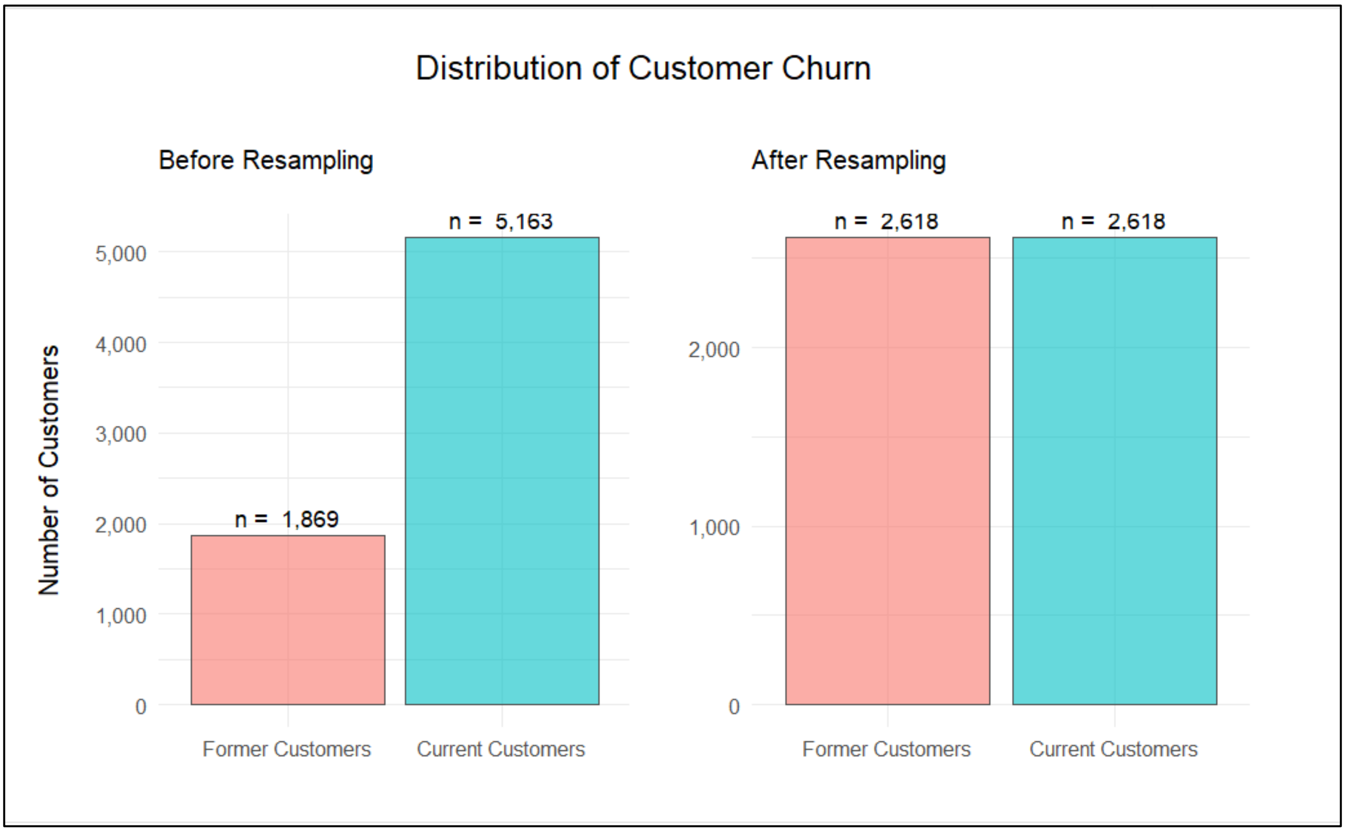
- Our target variable, Churn, is quite imbalanced with a little over 26% (1,869 customers) leaving the company within the past month. Since class imbalance can negatively affect the precision and recall accuracy of statistical models, I will use a synthetic minority over-sampling technique known as smote to create a more evenly distributed training set.
- The smote algorithm artificially generates new instances of the minority class using the nearest neighbors of these cases and under-samples the majority class to create a more balanced data set. After applying smote, our training set now consists of an equal proportion of current and former customers

Distribution of Customer Churn

```{r}
telecom <- telecom %>%
  mutate_at(15, ~as.factor(case_when(. == "One year" ~ "OneYear",
                                     . == "Two year" ~ "TwoYear",
                                     TRUE ~ "Month-to-month")))
set.seed(1)
ind <- createDataPartition(telecom$Churn, p = 0.7, list = F)
telecom.train <- telecom[ind,]
telecom.test <- telecom[-ind,]
train.resamp <- smote(Churn ~ ., data = data.frame(telecom.train), perc.over = 1,
perc.under = 2)
g1 <- ggplot(t2, aes(x = fct_rev(Churn2), fill = fct_rev(Churn2))) +
  geom_bar(alpha = 0.6, color = "grey30", show.legend = F) +
  geom_text(stat = "count", size = 3.5,
            aes(label = paste("n = ", formatC(..count.., big.mark = ","))), vjust = -0.5) +
  scale_y_continuous(labels = comma_format()) +
  labs(subtitle = "Before Resampling\n", x = "", y = "Number of Customers\n")


g2 <- ggplot(train.resamp, aes(x = fct_rev(ifelse(Churn == "Yes", "Former
Customers", "Current Customers")),
                               fill = fct_rev(Churn))) +
  geom_bar(alpha = 0.6, color = "grey30", show.legend = F) +
  geom_text(stat = "count", size = 3.5,
            aes(label = paste("n = ", formatC(..count.., big.mark = ","))), vjust = -0.5) +
  scale_y_continuous(labels = comma_format()) +
  labs(subtitle = "After Resampling\n", x = "", y = "")
options(repr.plot.width=9, repr.plot.height=7)
grid.arrange(g1, g2, nrow = 1, top = textGrob("Distribution of Customer Churn\n",
                                              gp = gpar(fontsize = 14)))
```

```




## 3.4 Feature Selection

To identify which features should be included in the models, I will use a two-step process. First, I will check the chi-squared tests of independence between the categorical features and include only variables that have a statistically significant association to our response, Churn. Then, I will use the random forest algorithm to identify the most important predictors of customer churn.

## Chi-Squared Tests

The Chi-Squared Test of Independence evaluates the association between two categorical variables. The null hypothesis for this test is that there is no relationship between our response variable and the categorical feature, and the alternative hypothesis is that that there is a relationship. Looking at the results of the tests, Gender and PhoneService have very small chi-squared statistics and p-values that are greater than the significance threshold, a, of 0.05, indicating they are independent of our target variable. The rest of the categorical features do have a statistically significant association to customer churn.
```

```{r}

# Perform Chi-Squared tests, excluding the target variable from the predictors
chi <- lapply(names(categorical_vars)[-17], function(col_name) {
  result <- chisq.test(categorical_vars[, 17], categorical_vars[[col_name]])
  result$variable <- col_name # Add the variable name
  result
})

# Convert Chi-Squared test results into a tidy data frame with variable names
chi_results <- do.call(rbind, lapply(chi, function(res) {
  tidy_res <- broom::tidy(res)
  tidy_res$variable <- res$variable # Add the variable name to the tidy results
  tidy_res
})) %>%
  arrange(p.value) %>%
  mutate(across(c(statistic, p.value), ~ round(., 3)))

# View the results
chi_results
```

	statistic	p.value	parameter	method	variable
1	493.524	0.000	2	Pearson's Chi-squared test	InternetService
2	351.922	0.000	2	Pearson's Chi-squared test	TechSupport
3	310.492	0.000	1	Pearson's Chi-squared test with Yates' continuity correction	Dependents
4	312.264	0.000	2	Pearson's Chi-squared test	OnlineSecurity
5	268.742	0.000	3	Pearson's Chi-squared test	PaymentMethod
6	250.366	0.000	2	Pearson's Chi-squared test	StreamingMovies
7	241.757	0.000	2	Pearson's Chi-squared test	StreamingTV
8	235.061	0.000	2	Pearson's Chi-squared test	DeviceProtection
9	234.328	0.000	2	Pearson's Chi-squared test	OnlineBackup
10	170.836	0.000	1	Pearson's Chi-squared test with Yates' continuity correction	PaperlessBilling
11	158.441	0.000	1	Pearson's Chi-squared test with Yates' continuity correction	Churn
12	151.397	0.000	2	Pearson's Chi-squared test	MultipleLines
13	144.127	0.000	2	Pearson's Chi-squared test	Contract
14	1.931	0.165	1	Pearson's Chi-squared test with Yates' continuity correction	Partner
15	0.421	0.516	1	Pearson's Chi-squared test with Yates' continuity correction	PhoneService
16	0.014	0.904	1	Pearson's Chi-squared test with Yates' continuity correction	gender

CHAPTER IV

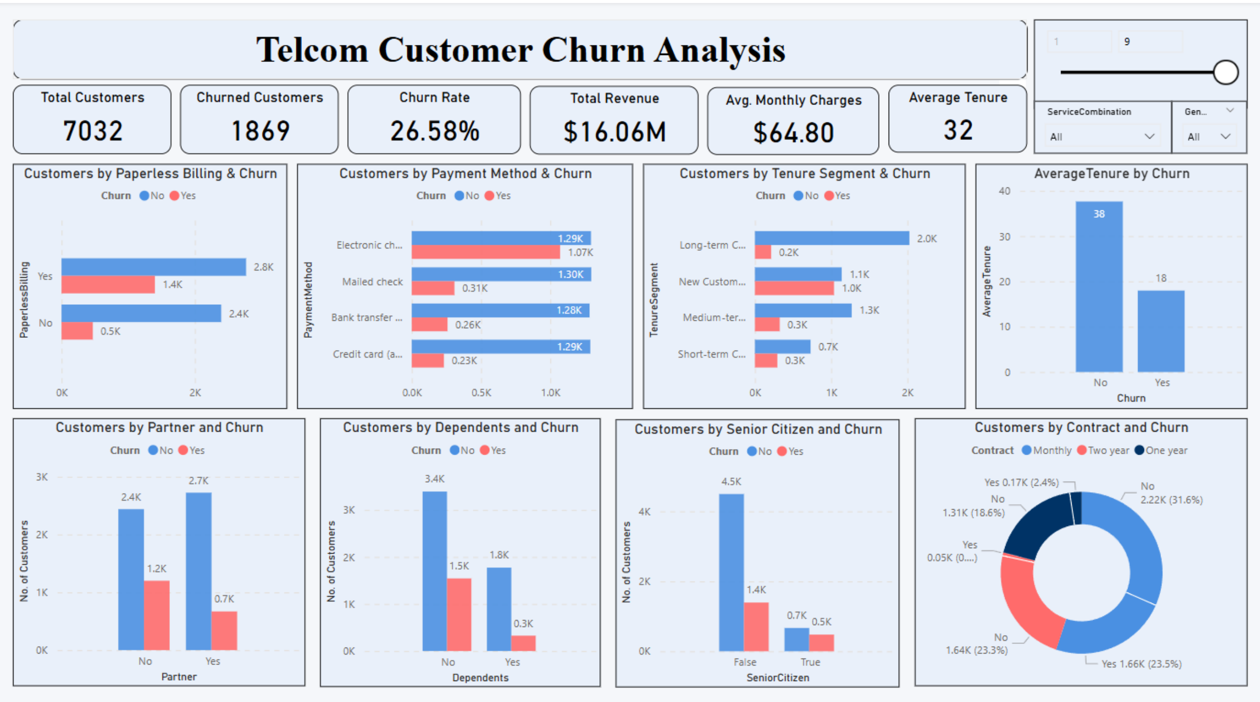
BUSINESS INTELLIGENCE INTERACTIVE DASHBOARDS

TELECOM CHURN ANALYSIS

4.1 Dashboards Interpretation

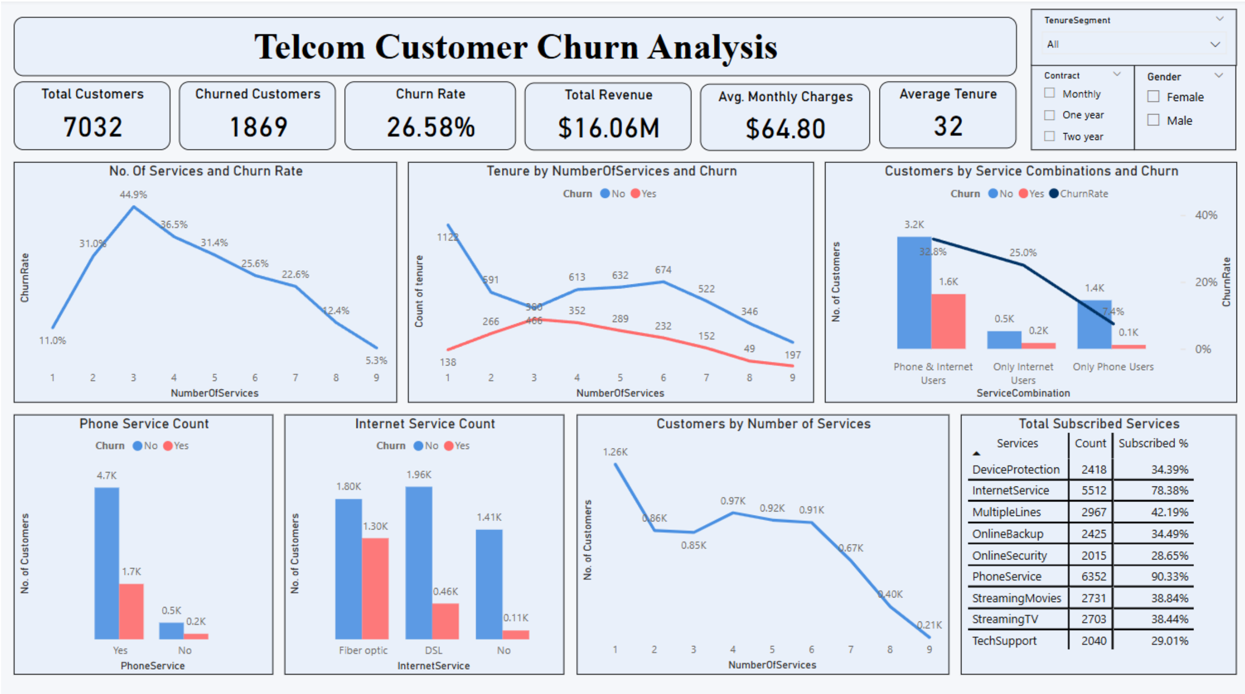
The interactive dashboard provides insights into telecom customer churn, displaying key metrics such as churn rate, revenue, average monthly charges, and customer tenure. This analysis helps in understanding the factors influencing customer churn and retention.

Telecom Customer Churn Analysis - Overview Dashboard



This chapter presents an in-depth analysis of the telecom churn dataset using interactive dashboards. Each chart provides insights into customer behavior, churn trends, and key influencing factors. Additionally, the DAX functions used to create these charts are included for reference.

Telecom Customer Churn Analysis - Service & Tenure Dashboard



4.2 Key Metrics Summary

- **Total Customers:** 7032
- **Churned Customers:** 1869
- **Churn Rate:** 26.58%
- **Total Revenue:** $16.06M
- **Average Monthly Charges:** $64.80
- **Average Tenure:** 32 months

These summary metrics offer a high-level perspective on the customer base and churn patterns.

Key DAX Measures Used

1.Average Customer Tenure

AverageTenure = AVERAGE('Telco-Customer-Churn'[tenure])

This measure calculates the average tenure of customers in months.

2.Average Monthly Charges

```
AvgMonthlyCharges = DIVIDE(
    SUM('Telco-Customer-Churn'[MonthlyCharges]),
    COUNTROWS('Telco-Customer-Churn')
)
```

This computes the average monthly charge per customer.

3. Total Churned Customers

```
ChurnedCustomers = CALCULATE(
    COUNTROWS('Telco-Customer-Churn'),
    'Telco-Customer-Churn'[Churn] = "Yes"
)
```

This measure determines the total number of customers who have churned.

4. Churn Rate

```
ChurnRate = DIVIDE('Telco-Customer-Churn'[ChurnedCustomers], COUNT('Telco-
        Customer-Churn'[customerID]))
```

This calculates the percentage of customers who have churned.

5. Number of Services Availed

```
NumberOfServices =
    IF([PhoneService] = "Yes", 1, 0) +
    IF([MultipleLines] = "Yes", 1, 0) +
    IF([InternetService] = "DSL" || [InternetService] = "Fibre Optic", 1, 0) +
    IF([OnlineSecurity] = "Yes", 1, 0) +
    IF([OnlineBackup] = "Yes", 1, 0) +
    IF([DeviceProtection] = "Yes", 1, 0) +
    IF([TechSupport] = "Yes", 1, 0) +
    IF([StreamingTV] = "Yes", 1, 0) +
    IF([StreamingMovies] = "Yes", 1, 0)
```

This measure counts the total number of services subscribed by a customer.

6. Retention Rate

```
RetentionRate =
CALCULATE(
    COUNTROWS('Telco-Customer-Churn'),
    'Telco-Customer-Churn'[Churn] = "No"
) / COUNTROWS('Telco-Customer-Churn')
```

This calculates the percentage of customers who are retained.

7. Revenue Lost Due to Churn

```
RevenueLost = CALCULATE(
    SUM('Telco-Customer-Churn'[TotalCharges]),
    'Telco-Customer-Churn'[Churn] = "Yes"
)
```

This measure computes the total revenue lost from churned customers.

8. Service Combination Categorization

```
ServiceCombination =
    SWITCH(
        TRUE(),
        'Telco-Customer-Churn'[PhoneService] = "Yes" && 'Telco-Customer-
           Churn'[InternetService] IN {"Fibre optic", "DSL"}, "Phone & Internet Users",
        'Telco-Customer-Churn'[PhoneService] = "Yes" && 'Telco-Customer-
           Churn'[InternetService] = "No", "Only Phone Users",
        'Telco-Customer-Churn'[PhoneService] = "No" && 'Telco-Customer-
           Churn'[InternetService] IN {"Fibre optic", "DSL"}, "Only Internet Users",
        "Unknown"
    )
```

This categorizes customers based on the services they subscribe to.

9. Tenure Segmentation

```
TenureSegment =
    SWITCH(
        TRUE(),
        [Tenure] >= 1 && [Tenure] <= 12, "New Customer",
        [Tenure] >= 13 && [Tenure] <= 24, "Short-term Customer",
        [Tenure] >= 25 && [Tenure] <= 48, "Medium-term Customer",
        [Tenure] >= 49 && [Tenure] <= 72, "Long-term Customer",
        "Unknown"
    )
```

This segments customers into different tenure groups based on their subscription duration.

10. Total Customers

```
TotalCustomers = CALCULATE(COUNTROWS('Telco-Customer-Churn'))
```

This calculates the total number of customers in the dataset.

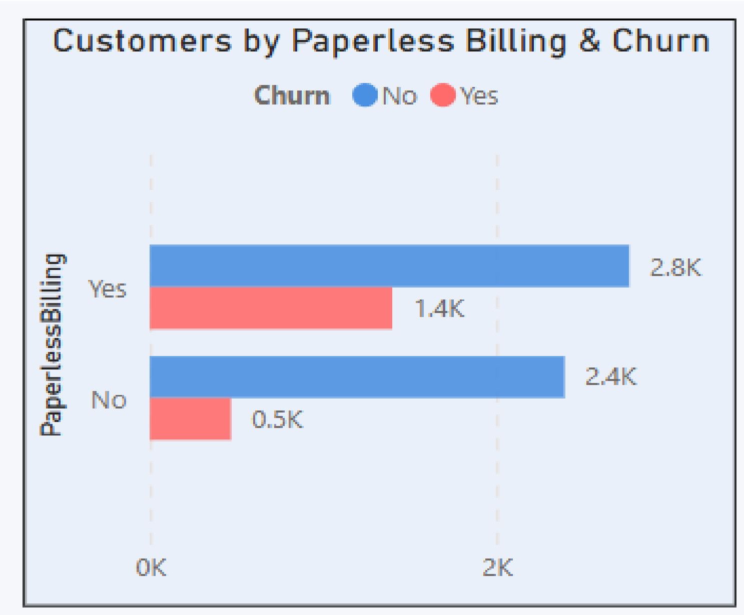
11. Total Revenue

```
TotalRevenue = SUM('Telco-Customer-Churn'[TotalCharges])
```

This computes the total revenue earned from customers.

4.2 Chart Insights

4.2.1 Insight: Customers by Paperless Billing & Churn



The chart above displays the distribution of customers based on their choice of **Paperless Billing** and their corresponding **Churn status**.

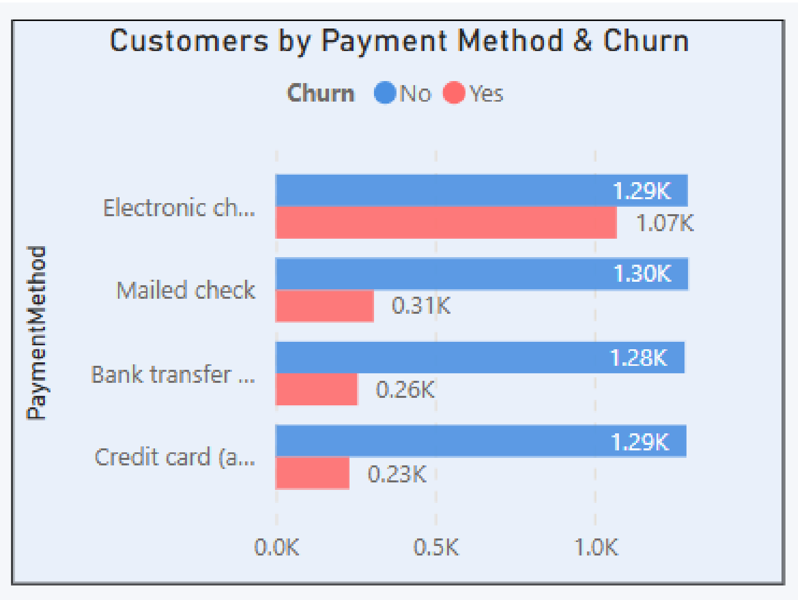
Key Observations:

- Customers who have opted for **Paperless Billing** have a significantly higher churn rate.
  - Out of **2.8K customers** with Paperless Billing, **1.4K** (50%) have churned.
- In contrast, customers who do **not** use Paperless Billing show **lower churn rates**.
  - Out of **2.4K customers** without Paperless Billing, only **0.5K** (≈21%) have churned.
- This suggests that Paperless Billing might be associated with a **higher likelihood of customer churn**.

Potential Business Insights:

- The higher churn rate among Paperless Billing users may indicate **customer dissatisfaction** with online billing or associated services.
- Businesses could **investigate user feedback** to understand issues related to Paperless Billing, such as billing transparency, ease of payment, or customer service concerns.
- Offering **incentives**, improved UI/UX for online billing, or **personalized support** for Paperless Billing users may help reduce churn.

4.2.2 Insight: Customers by Payment Method & Churn



The chart above illustrates the distribution of customers based on their **Payment Method** and their corresponding **Churn status**.

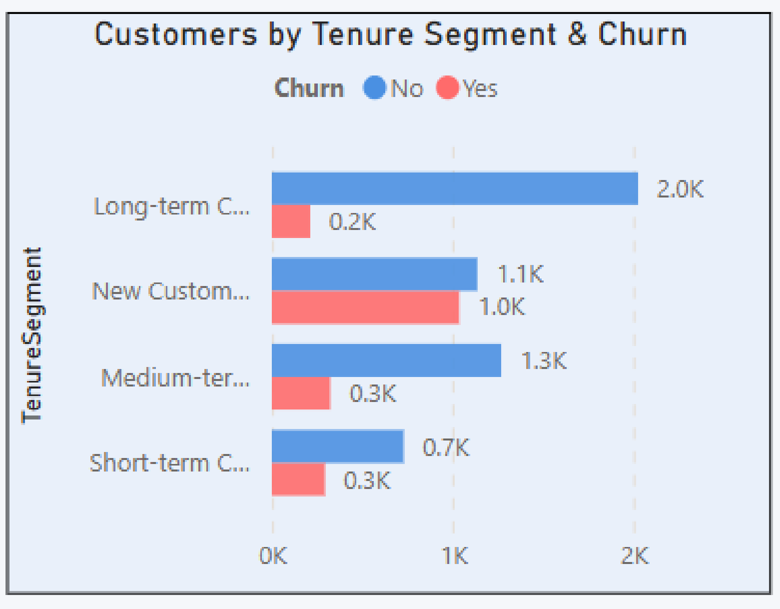
Key Observations:

- **Electronic Check users have the highest churn rate** compared to other payment methods.
    - Out of **2.36K** Electronic Check users, **1.07K (≈45%)** have churned.
- **Mailed Check, Bank Transfer, and Credit Card users show significantly lower churn rates.**
    - Mailed Check: **0.31K churned** out of **1.61K total users (≈19%)**
    - Bank Transfer: **0.26K churned** out of **1.54K total users (≈17%)**
    - Credit Card: **0.23K churned** out of **1.52K total users (≈15%)**

Potential Business Insights:

- Customers paying via **Electronic Check** exhibit **the highest churn rate**, possibly due to **transaction failures, security concerns, or inconvenience.**
- Customers using **Credit Cards and Bank Transfers are more likely to stay**, indicating that these methods may be more convenient or reliable.
- **Actionable Strategy:**
    - Encourage users to switch from **Electronic Check to more stable payment options** like Credit Card or Bank Transfer.
    - Offer **incentives or discounts** for customers who switch to a more reliable payment method.
    - Investigate **customer feedback** on issues with Electronic Check payments to reduce churn.

4.2.3 Insight: Customers by Tenure Segment & Churn



The chart above presents customer distribution across different **tenure segments** and their **churn behavior**.

Key Observations:

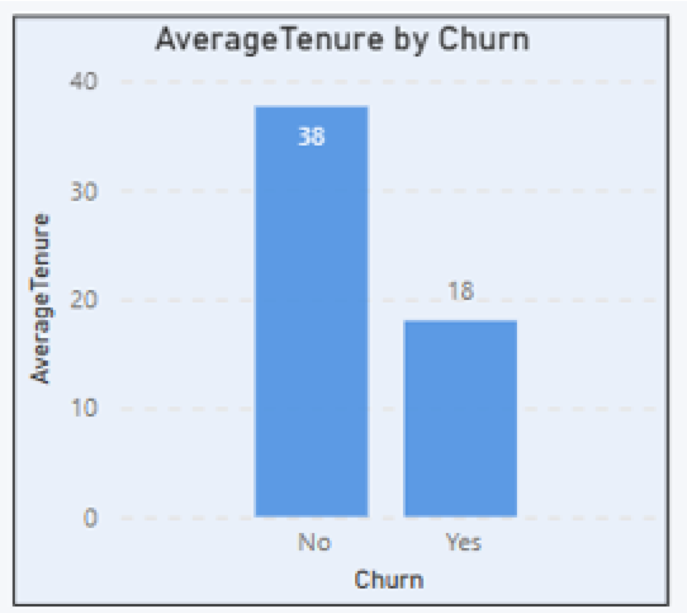
- **New Customers (1–12 months) have the highest churn rate**, with **1.0K churned out of 2.1K total (≈48%)**.
- **Short-term (13–24 months) and Medium-term (25–48 months) customers** exhibit moderate churn rates of **~30%**.
- **Long-term customers (49+ months) have the lowest churn rate**, with only **0.2K churned out of 2.2K total (≈9%)**.

Potential Business Insights:

- **High churn among New Customers suggests dissatisfaction early in the customer journey.**
- **Customers who stay beyond 2 years (Medium & Long-term) are significantly more loyal.**
- **Actionable Strategy:**
  - Enhance **onboarding experience** and offer better initial incentives to **retain new customers**.

- Implement **early engagement strategies** like personalized discounts, better customer support, and loyalty rewards.
- Conduct **exit surveys** for new customers who churn to identify key dissatisfaction factors.

4.2.4 Insight: Average Tenure by Churn



The bar chart above displays the average tenure of customers who churned versus those who remained subscribed.

Key Observations:

- Customers who did **not churn** have an average tenure of **38 months**.
- Customers who **churned** have a significantly lower average tenure of **18 months**.

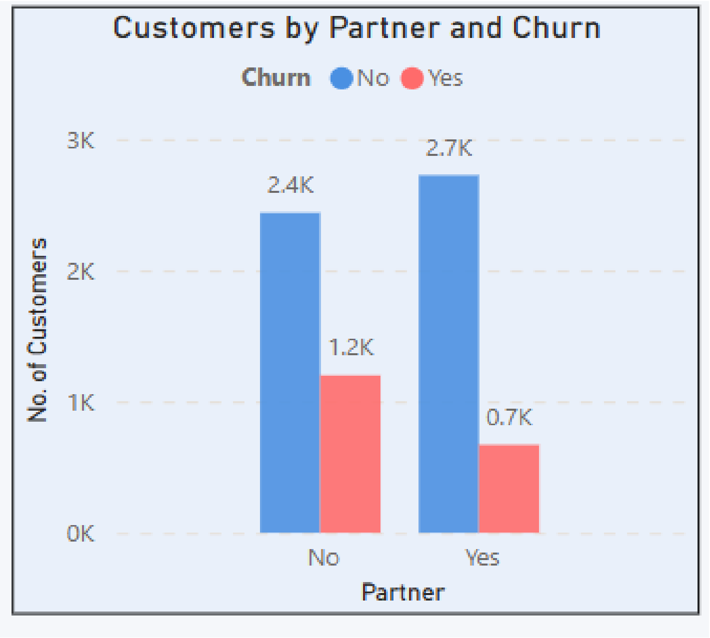
Potential Business Insights:

- Customers who **stay longer (38 months)** are more likely to remain loyal, indicating that long-term users are satisfied with the service.
- Churned customers tend to leave early, suggesting a **critical retention period** in the first 18 months.

Actionable Strategy:

- **Early Engagement:** Implement personalized offers, better onboarding, and proactive support within the first **18 months** to reduce churn risk.
- **Retention Campaigns:** Target medium-term customers (~18 months) with incentives such as loyalty bonuses or service upgrades to encourage long-term commitment.
- **Churn Prediction Model:** Use **machine learning models** to identify customers likely to churn within 18 months and take preventive actions.

4.2.5 Insight: Customers by Partner and Churn



The bar chart above illustrates the distribution of customers based on their partner status and churn behavior.

Key Observations:

- Customers **with a partner** have a lower churn rate: **0.7K churned out of 3.4K total (~20.6%)**.
- Customers **without a partner** have a significantly higher churn rate: **1.2K churned out of 3.6K total (~33.3%)**.
- The presence of a partner appears to correlate with greater customer retention.

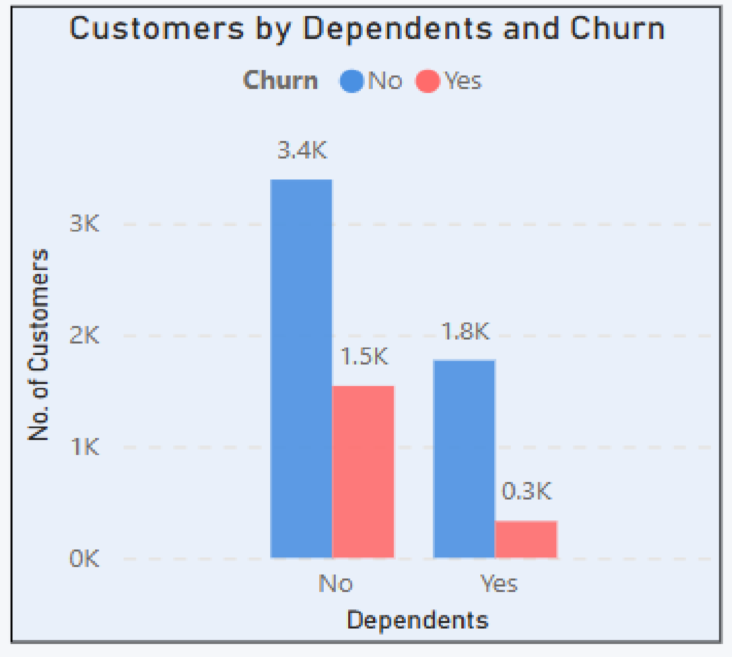
Potential Business Insights:

- Customers **without a partner** might lack a shared household commitment, making them more likely to switch or cancel.
- Customers **with a partner** may have more stable usage patterns, possibly due to shared household consumption.

Actionable Strategy:

- **Targeted Retention Offers:** Provide personalized plans or family/partner discounts to encourage long-term subscriptions.
- **Engagement Strategies:** Introduce referral programs to incentivize customers without a partner to bring in a secondary user, enhancing retention.
- **Churn Reduction Campaigns:** Offer bundled services or loyalty perks tailored for single users to boost their long-term engagement.

4.2.6 Insight: Customers by Dependents and Churn



The bar chart above illustrates customer distribution based on dependent status and their churn behavior.

Key Observations:

- Customers **without dependents** have a higher churn rate: **1.5K churned out of 4.9K total (~30.6%)**.
- Customers **with dependents** have a significantly lower churn rate: **0.3K churned out of 2.1K total (~14.3%)**.
- Having dependents appears to be associated with higher customer retention.

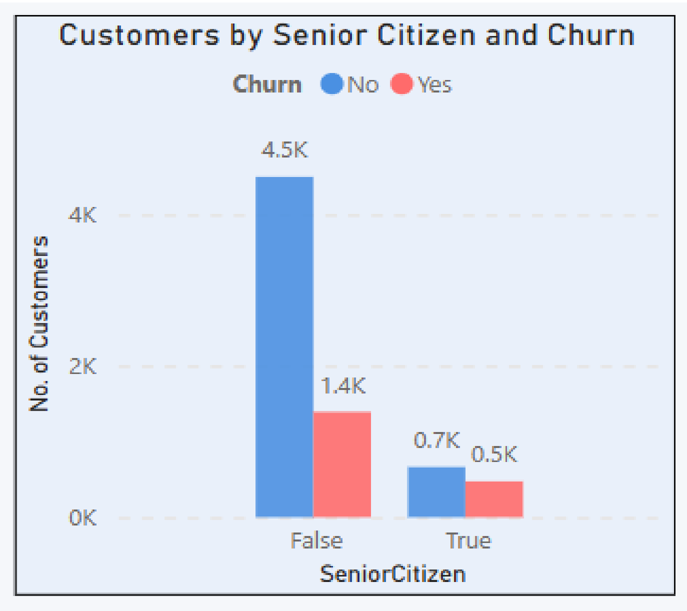
Potential Business Insights:

- Customers **without dependents** may have more flexibility in switching providers, leading to higher churn.
- Customers **with dependents** might seek **stability and continuity**, making them less likely to churn.

Actionable Strategy:

- **Family-Oriented Plans:** Offer exclusive family bundles or multi-user discounts to appeal to customers with dependents.
- **Loyalty Programs:** Provide long-term benefits for users with dependents to enhance retention.
- **Targeted Engagement:** Design campaigns addressing the flexibility needs of customers without dependents, such as **customizable plans or premium content access**.

4.2.7 Insight: Customers by Senior Citizen and Churn



The bar chart above displays customer distribution based on senior citizen status and their churn behavior.

Key Observations:

- **Non-senior customers (False)** have a churn rate of **1.4K out of 5.9K total (~23.7%)**.
- **Senior citizens (True)** have a significantly higher churn rate of **0.5K out of 1.2K total (~41.7%)**.
- Senior citizens churn at a much higher rate compared to younger customers.

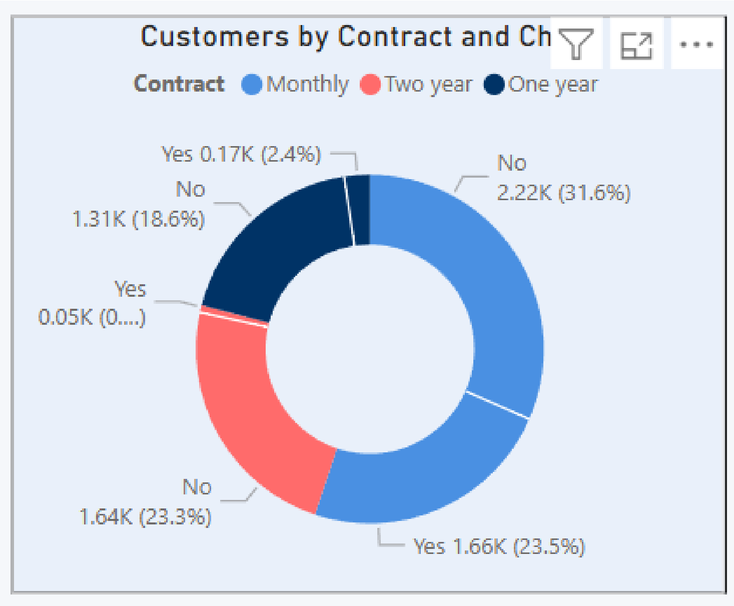
Potential Business Insights:

- Senior citizens may find the service **less user-friendly**, leading to dissatisfaction.
- They might have different usage preferences, requiring **tailored offerings**.
- Financial concerns (fixed income, budget-conscious behavior) could influence their decision to churn.

Actionable Strategy:

- **Simplified User Experience:** Offer an easy-to-use interface, larger text, and simplified navigation for senior customers.
- **Exclusive Senior Plans:** Provide discounted senior-friendly plans with essential features at an affordable price.
- **Enhanced Customer Support:** Introduce **dedicated support channels** (hotline, chat assistance) tailored to senior users.
- **Educational Workshops:** Offer free tutorials on how to maximize service benefits to improve engagement and retention.

4.2.8 Insight: Customers by Contract Type and Churn



The donut chart above illustrates customer distribution based on contract type and churn behavior.

Key Observations:

- **Monthly contract customers** have the highest churn rate, with **1.31K churned out of 1.48K total (~88.5%)**.
- **One-year contract customers** exhibit a moderate churn rate, with **0.17K churned out of 1.83K total (~9.3%)**.
- **Two-year contract customers** have the lowest churn rate, with only **0.05K churned out of 1.69K total (~3.0%)**.

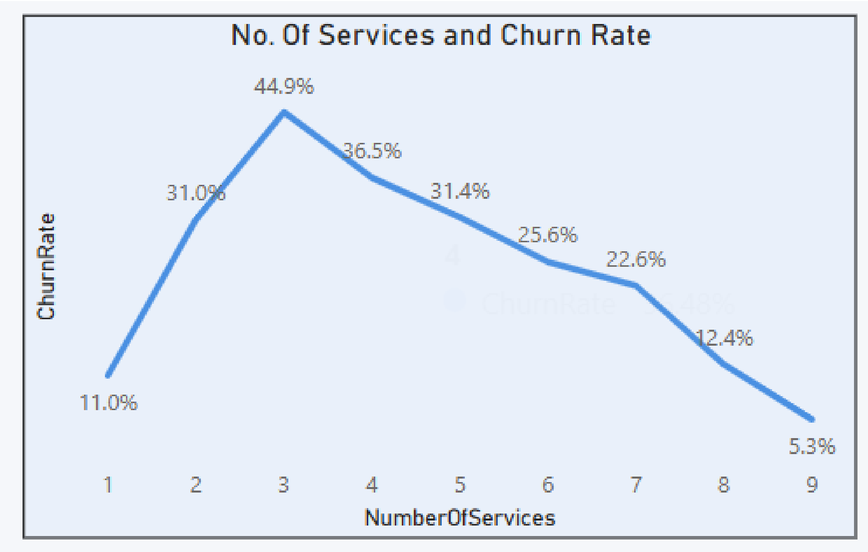
Potential Business Insights:

- Customers on **monthly contracts** are more likely to leave, likely due to the flexibility to cancel at any time.
- **Longer contract durations (one-year and two-year)** significantly reduce churn, suggesting that commitment-based plans improve retention.
- Customers with **longer contracts** may find the service valuable enough to commit long-term, or they may be incentivized by better pricing.

Actionable Strategy:

- **Incentivize Longer Commitments:** Offer discounts, additional features, or exclusive perks for customers who switch from monthly to annual plans.
- **Improve Monthly Customer Retention:** Provide loyalty rewards or flexible upgrade options to encourage them to transition to longer contracts.
- **Identify At-Risk Monthly Users:** Use predictive churn models to detect customers likely to leave and engage them with personalized retention offers.

4.2.9 Insight: Number of Services and Churn Rate



The line chart above shows the relationship between the number of services subscribed to and the corresponding churn rate.

Key Observations:

- Customers subscribing to **only 1 service** have a relatively low churn rate of **11.0%**.
- Churn rate peaks at **44.9% for customers with 3 services**, indicating higher dissatisfaction or complexity at this level.
- Beyond 3 services, churn rate **gradually declines**, reaching just **5.3% for customers with 9 services**.
- Customers with **7+ services are significantly less likely to churn**, suggesting strong engagement and satisfaction.

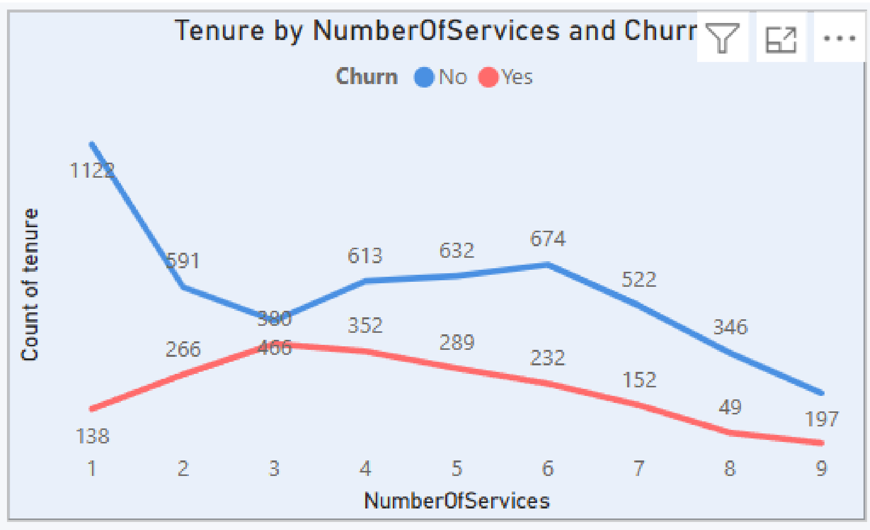
Potential Business Insights:

- Customers with **fewer services (1-3)** may not see enough value, leading to higher churn.
- A **sweet spot for engagement** appears at **4+ services**, where churn stabilizes and then drops.
- Customers with **multiple services (7+)** are highly retained, possibly due to service bundling or better perceived value.

Actionable Strategy:

- **Encourage Service Bundling:** Promote **discounted multi-service packages** to increase customer retention.
- **Target 3-Service Subscribers:** Identify pain points through surveys and offer personalized incentives to prevent churn.
- **Upsell & Cross-Sell Strategies:** Encourage customers with 1-2 services to try additional offerings through trial promotions.
- **Monitor Complexity:** Ensure that adding more services does not overwhelm customers, leading to dissatisfaction.

4.2.10 Insight: Tenure by Number of Services and Churn



The line chart above visualizes the relationship between the **number of services subscribed to**, **tenure**, and **churn behavior**.

Key Observations:

- Customers with **1 service** have the lowest tenure and a high churn count, indicating weaker engagement.
- As the **number of services increases**, tenure **increases for non-churned customers**, showing that multi-service users tend to stay longer.
- The churned customer count **peaks at 3 services** and then declines, suggesting that customers with 3 services may experience dissatisfaction or complexity.
- Customers with **6+ services tend to stay longer and churn less**, reinforcing the trend observed in the previous analysis.

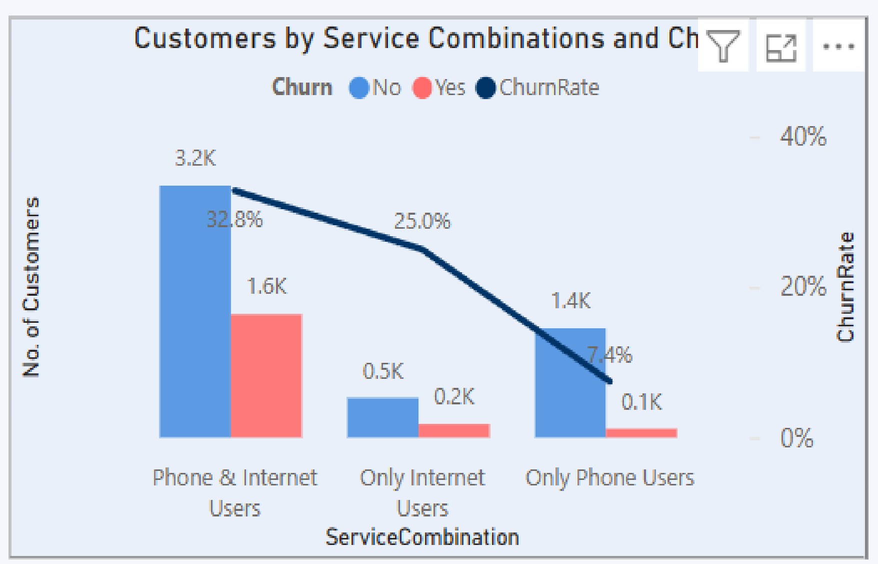
Potential Business Insights:

- **Single-service users have shorter tenures**, indicating that they may not find enough value to stay long-term.
- Customers who subscribe to **multiple services (4+) tend to remain loyal** and have longer tenures.
- The **drop in tenure for churned customers beyond 3 services** suggests a threshold where service complexity or cost may be a deciding factor for leaving.

Actionable Strategy:

- **Improve Value for Single-Service Users:** Offer bundled promotions or highlight benefits of additional services to encourage multi-service adoption.
- **Identify High-Risk 3-Service Customers:** Use predictive churn models to detect dissatisfaction and provide personalized retention incentives.
- **Enhance Customer Support & Onboarding:** Ensure that customers who subscribe to multiple services receive adequate support to prevent churn due to complexity.
- **Offer Long-Term Benefits:** Reward customers who commit to multiple services for longer durations to enhance engagement and reduce churn risk.

4.2.11 Insight: Customers by Service Combinations and Churn



The chart above illustrates churn behavior across different service combinations, highlighting the **number of customers** and **churn rate** for each category.

Key Observations:

- **Phone & Internet Users** have the **highest churn rate (32.8%)**, with **1.6K out of 4.8K** customers leaving.
- **Only Internet Users** show a **moderate churn rate (25.0%)**, with **0.2K churned out of 0.8K total customers**.
- **Only Phone Users** experience the lowest churn rate at **7.4%**, with **only 0.1K churned out of 1.5K customers**.
- The overall trend suggests that **bundled service users (Phone & Internet) are more prone to churn** than those using a single service.

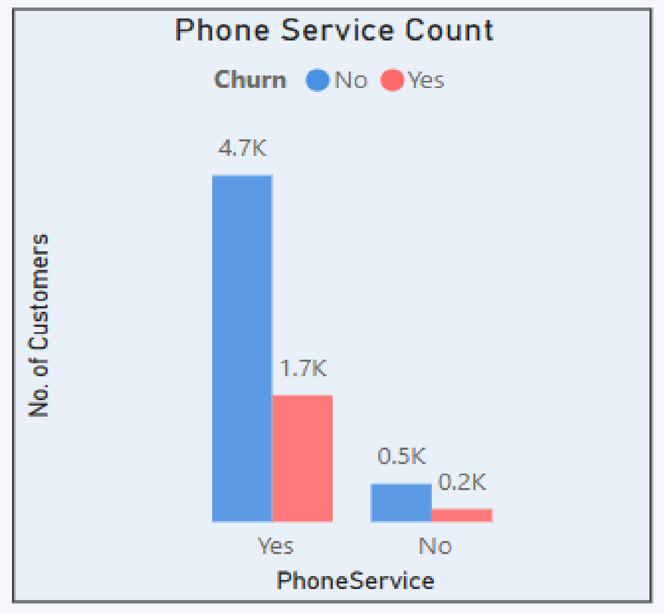
Potential Business Insights:

- **Bundled users have the highest churn**, potentially due to higher costs, service dissatisfaction, or complex billing issues.
- **Internet-only users also experience a moderate churn rate**, indicating that standalone internet services may not offer enough value.
- **Phone-only users are the most stable**, possibly due to long-term contracts or lower dependency on additional services.

Actionable Strategy:

- **Enhance Value for Bundled Users:** Offer exclusive benefits like discounts, priority support, or added features to retain them.
- **Investigate Internet Service Issues:** Conduct customer feedback surveys to identify pain points among **internet-only users**.
- **Upsell to Phone-Only Users:** Encourage phone-only users to explore internet services through **trial offers or limited-time discounts**.
- **Simplify Billing & Plans:** Ensure that bundled customers do not face unexpected charges or service complexities that might drive them away.

4.2.12 Insight: Phone Service and Churn



The chart above presents churn behavior based on whether customers have **phone service** or not.

Key Observations:

- **Customers with Phone Service**:
  - **4.7K customers retained** (No Churn).
  - **1.7K customers churned**, resulting in a churn rate of **≈26.6%**.
- **Customers without Phone Service**:
  - **0.5K customers retained**.
  - **0.2K customers churned**, leading to a slightly higher churn rate of **≈28.6%**.

- Customers without phone service appear to have a slightly **higher churn rate than those with phone service**.

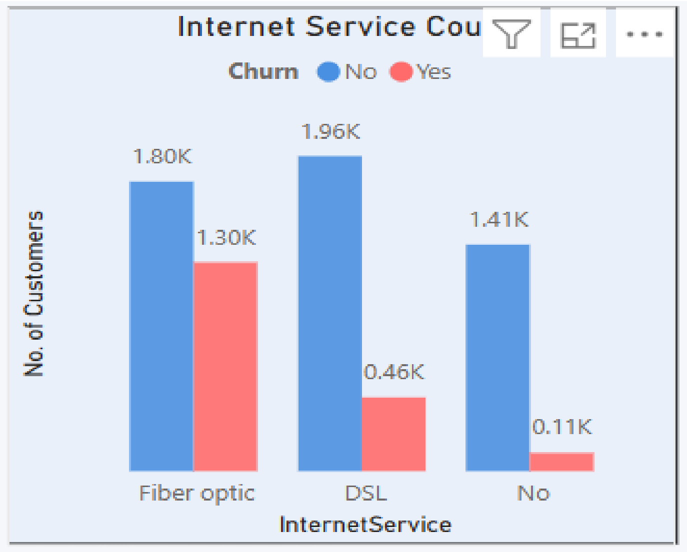
Potential Business Insights:

- Having a **phone service does not significantly reduce churn**, suggesting that customers may not view phone services as a strong retention factor.
- Customers **without phone service may be less engaged with the company**, leading to higher churn.
- There is still a **substantial number of phone service users churning**, indicating dissatisfaction in other areas like pricing, network quality, or bundled service experiences.

Actionable Strategy:

- **Investigate Reasons for Churn**: Conduct surveys among phone and non-phone users to understand their motivations for leaving.
- **Improve Phone Service Value**: Offer better **bundled deals, enhanced call quality, or exclusive perks** for phone service users.
- **Upsell Phone Services to Non-Users**: Offer promotions, **free trials, or discounts** to encourage more users to opt for phone services.
- **Cross-Sell Additional Services**: Customers with only **internet or other services** may benefit from bundled offers that include **phone services at a discounted rate**.

4.2.13 Insight: Internet Service and Churn



The chart above displays the churn behaviour among customers based on their type of **Internet Service**.

Key Observations:

- **Fibre Optic Users:**
  - **1.8K customers retained**, **1.3K churned** (**≈41.9% churn rate**).
- **DSL Users:**
  - **1.96K customers retained**, **0.46K churned** (**≈19% churn rate**).
- **Customers Without Internet Service:**
  - **1.41K customers retained**, **0.11K churned** (**≈7.2% churn rate**).
- **Fibre optic users have the highest churn rate**, whereas customers without internet service have the lowest churn rate.

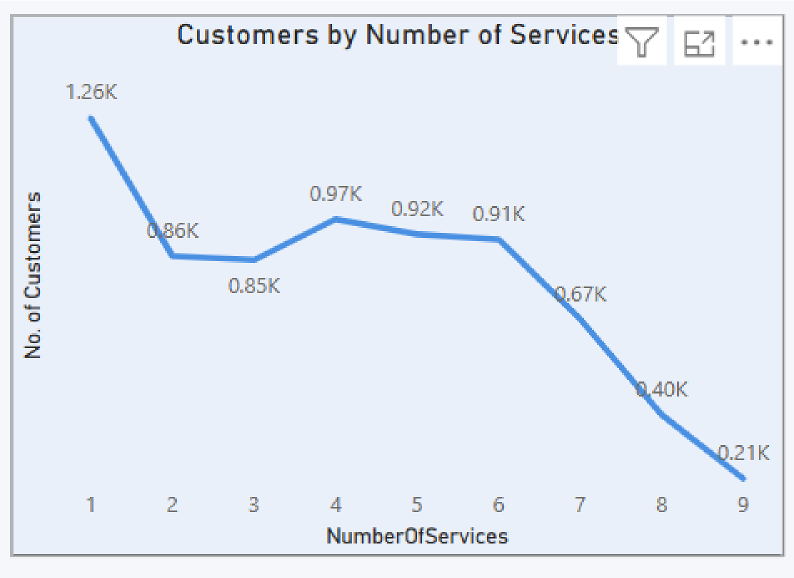
Potential Business Insights:

- **Fibre optic service might be causing dissatisfaction** due to **pricing, service quality, or competition**.
- **DSL users have significantly lower churn rates**, indicating **higher satisfaction or lower pricing**.
- Customers **without internet service have the lowest churn rate**, suggesting they may be **more reliant on other services** (e.g., phone).

Actionable Strategy:

- **Investigate Fibre Optic Issues**: Conduct customer feedback surveys to identify dissatisfaction points (e.g., **speed, outages, cost**).
- **Competitive Pricing for Fibre Optic**: Offer **discounts or promotional deals** to retain fibre optic users.
- **Encourage DSL Users to Upgrade**: Provide **upgrade incentives** to DSL users for a smoother transition to fibre optic.
- **Cross-Sell Internet Services to Non-Users**: Since non-internet users have low churn, **target them with internet bundle offers** to improve retention.

4.2.14 Insight: Customer Distribution by Number of Services



The above chart illustrates the **number of customers** subscribed to different counts of telecom services.

Key Observations:

- The **highest number of customers (1.26K) subscribe to only 1 service**.
- The customer count **drops sharply** after 3 services, with a **steady decline** as the number of services increases.
- Only **210 customers subscribe to all 9 services**, indicating **low adoption of full-service bundles**.

Potential Business Insights:

- Customers prefer **fewer services**, likely due to **pricing, lack of perceived value, or specific needs**.
- The **drop after 3 services** suggests that **bundling more than three services might not be attractive** to most customers.
- **Limited adoption of high-tier packages (7+ services)** could indicate a need for **better marketing, pricing adjustments, or improved service benefits**.

Actionable Strategy:

- **Bundle Optimization:** Create attractive **3-service bundles** based on customer needs to **increase adoption**.
- **Promotional Offers for High-Tier Packages:** Provide **discounts or added benefits** to incentivize users to **upgrade beyond 3 services**.
- **Customer Segmentation Analysis:** Identify which services are most commonly paired and target users with personalized **service recommendations**.
- **Feedback Collection:** Conduct surveys to understand **why customers avoid high-tier service bundles** and **opt for limited services**.

4.2.15 Insight: Subscription Distribution Across Services

Total Subscribed Services

Services	Count	Subscribed %
DeviceProtection	2418	34.39%
InternetService	5512	78.38%
MultipleLines	2967	42.19%
OnlineBackup	2425	34.49%
OnlineSecurity	2015	28.65%
PhoneService	6352	90.33%
StreamingMovies	2731	38.84%
StreamingTV	2703	38.44%
TechSupport	2040	29.01%

The above table provides a **breakdown of customer subscriptions** across various telecom services, along with their adoption percentages.

Key Observations:

- **Phone Service (90.33%) and Internet Service (78.38%)** have the highest adoption rates, indicating they are **core services** for most customers.
- **Multiple Lines (42.19%)** is relatively common, suggesting many users **opt for multiple connections**.
- **Value-Added Services (VAS) Adoption is Low:**

- **Device Protection (34.39%)**, **Online Backup (34.49%)**, and **Streaming Services (~38%)** show **moderate adoption**.
- **Online Security (28.65%)** and **Tech Support (29.01%)** have the **lowest uptake**, indicating a potential **gap in perceived value or awareness**.

Potential Business Insights:

- **Bundling Opportunities:**
  - Since **Internet Service is widely adopted**, bundling it with **Online Security or Tech Support** could **increase adoption**.
  - **Streaming Services (~38%)** can be marketed alongside **Internet Plans** to boost engagement.
- **Marketing & Awareness:**
  - Customers may **not fully understand** the benefits of **Online Security and Device Protection**, requiring better **education and targeted promotions**.
  - A **free trial period** could encourage more users to experience **Tech Support and Online Security** before committing.
- **Revenue Growth Strategy:**
  - Increase **cross-sell and upsell** strategies by promoting **discounted service bundles** (e.g., "Internet + Streaming TV + Tech Support").
  - Offer **family plans** for **Multiple Lines** to attract more **multi-user households**.

4.3 Recommendations

- **Offer discounted long-term contracts to reduce churn.**
- **Improve customer on boarding and engagement strategies for new users.**
- **Analyse and improve service quality for fibre optic users.**
- **Promote bundled service packages to retain customers.**

CHAPTER V

MODEL BUILDING

TELECOM CHURN ANALYSIS

5.1 Algorithm

To predict which customers are most likely to churn, several different types of classification models will be evaluated, including logistic regression, support vector machines, and random forests, KNN. Since the numeric predictors, MonthlyCharges and Tenure, have skewed distributions and varying scales, I will apply a preprocessing technique that normalizes the features to have a mean of 0 and a standard deviation of

5.2 Training and test dataset

10-fold cross-validation method for model building

What it does: Splits data into 70%-30% Train-Test Split:

70% Training Set: To train the model.

30% Test Set: To evaluate performance on unseen data.

10-Fold Cross-Validation

What it does:

Splits data into 10 parts (folds).

In each iteration:

9 folds = Training set.

1 fold = Test set.

Repeats 10 times, using each fold as the test set once.

Final result = Average of all 10 iterations.

Purpose: More reliable evaluation by minimizing randomness.

Advantages:

Uses all data for training and testing.

Provides more stable and accurate results.

Disadvantages:

Slower due to multiple training/testing cycles.

Slightly more complex than a single train-test split.

When to Use this method?

10-Fold Cross-Validation:

Best for small datasets where every point matters.

Preferred for thorough and reliable evaluations.

To fit the models, 10-fold cross-validation will be used and the model will be tested on the out of sample dataset. This set was held out of resampling and is more representative of the true class distribution.

5.3 Model

5.3.1 Logistic Regression

Logistic regression is a parametric classification technique that estimates the probability of an event occurring, for instance, whether or not a customer will leave the company. One of the advantages of the logistic model is the interpretability of the model parameters. Based on the size of the coefficients and the significance of the predictors, the model is able to quantify the relationships between our response and the input features.

````
```{r}

set.seed(21)

ctrl <- trainControl(method = "cv", number = 10, classProbs = TRUE,

                     summaryFunction = twoClassSummary)


glm.fit <- train(Churn ~ tenure + MonthlyCharges + InternetService + PaymentMethod +

                 Contract + OnlineSecurity + TechSupport + PaperlessBilling,

                 data = train.resamp, method = "glm", metric = "ROC",

                 preProcess = c("center", "scale"), trControl = ctrl)

glm.preds <- glm.fit %>% predict(telecom.test)

glm.cm <- data.frame(Logistic=confusionMatrix(glm.preds, telecom.test$Churn,

                                              positive = "Yes", mode = "everything")$byClass)
````
````

confusionMatrix(glm.preds, telecom.test$Churn, positive = "Yes", mode = "everything")

```
```

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1153  108
       Yes  395  452

               Accuracy : 0.7614
                 95% CI : (0.7426, 0.7794)
    No Information Rate : 0.7343
    P-Value [Acc > NIR] : 0.002448

                  Kappa : 0.4744

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8071
            Specificity : 0.7448
         Pos Pred Value : 0.5336
         Neg Pred Value : 0.9144
              Precision : 0.5336
                 Recall : 0.8071
                     F1 : 0.6425
             Prevalence : 0.2657
         Detection Rate : 0.2144
   Detection Prevalence : 0.4018
      Balanced Accuracy : 0.7760

       'Positive' Class : Yes
```

Our logistic regression model has an overall accuracy of **76.1%** and a precision of 53.3% on the test set. This means that when the model predicts a customer will leave, it is correct around 54% of the time. The recall of our model is 80.7%, which means that it correctly identified about 81% of all customers who left.

5.3.2 Support Vector Machine

Support vector machines (SVMs) are a commonly used statistical learning model. It is nonparametric, which means that it does not make any assumptions about the data like logistic regression does. SVMs involve finding a hyperplane that separates the data as well as possible and maximizes the distance between the classes of our response variable.

```
```{r}

svm.fit <- train(Churn ~ tenure + MonthlyCharges + InternetService + PaymentMethod
+

        Contract + OnlineSecurity + TechSupport + PaperlessBilling,

        data = train.resamp, method = "svmLinear", metric = "ROC",

        preProcess = c("center","scale"), trControl = ctrl)

svm.preds <- svm.fit %>% predict(telecom.test)

svm.cm <- data.frame(SVM=confusionMatrix(svm.preds, telecom.test$Churn,

                        positive = "Yes", mode = "everything")$byClass)

confusionMatrix(svm.preds, telecom.test$Churn, positive = "Yes", mode = "everything")
```
```

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  999  84
       Yes 549 476

               Accuracy : 0.6997
                 95% CI : (0.6796, 0.7192)
    No Information Rate : 0.7343
    P-Value [Acc > NIR] : 0.9998

                  Kappa : 0.3916

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8500
            Specificity : 0.6453
         Pos Pred Value : 0.4644
         Neg Pred Value : 0.9224
              Precision : 0.4644
                 Recall : 0.8500
                     F1 : 0.6006
             Prevalence : 0.2657
         Detection Rate : 0.2258
   Detection Prevalence : 0.4862
      Balanced Accuracy : 0.7477

       'Positive' Class : Yes
```

The accuracy of the linear support vector machine is about **69.9**% and the precision is 46%, which is not an improvement from the previous models. The recall did increase to 85%, which is the highest so far.

5.3.3 Random Forest

Random forest is a commonly used ensemble technique in machine learning. The model is built using a combination of many decision trees, where each takes a random sample of the data with replacement and selects a random subset of predictors, resulting in a relatively uncorrelated set of decision trees. Each tree then makes a prediction and the class with the most votes becomes the model's final prediction.

````
```{r}
rf.fit <- train(Churn ~ tenure + MonthlyCharges + InternetService + PaymentMethod +
                Contract + OnlineSecurity + TechSupport + PaperlessBilling,
                data = train.resamp, method = "rf", metric = "ROC",
                preProcess = c("center","scale"), trControl = ctrl)
rf.preds <- rf.fit %>%
  predict(telecom.test)
rf.cm <- data.frame(rf=confusionMatrix(rf.preds, telecom.test$Churn,
                                       positive = "Yes", mode = "everything")$byClass)
confusionMatrix(rf.preds, telecom.test$Churn, positive = "Yes", mode = "everything")
```
````

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1118  118
       Yes  430  442
                                          
               Accuracy : 0.74            
                 95% CI : (0.7208, 0.7587)
    No Information Rate : 0.7343          
    P-Value [Acc > NIR] : 0.2862          
                                          
                  Kappa : 0.4343          
                                          
 Mcnemar's Test P-Value : <2e-16          
                                          
            Sensitivity : 0.7893          
            Specificity : 0.7222          
         Pos Pred Value : 0.5069          
         Neg Pred Value : 0.9045          
              Precision : 0.5069          
                 Recall : 0.7893          
                     F1 : 0.6173          
             Prevalence : 0.2657          
         Detection Rate : 0.2097          
   Detection Prevalence : 0.4137          
      Balanced Accuracy : 0.7558          
                                          
       'Positive' Class : Yes             
```

The random forest classifier has an accuracy of **74%** and a precision of 50%, higher than the SVM but just below our logistic model. The recall of the model is about 75%, the lowest overall.

5.3.4 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a simple, non-parametric machine learning algorithm used for classification and regression tasks. It predicts the outcome based on the majority class or average value of its k nearest neighbors in the feature space. KNN is sensitive to the choice of k (the number of neighbors) and the scaling of the input features.

````
```{r}
#Train KNN model
set.seed(123)  # Set seed for reproducibility
knn.fit <- train(
  Churn ~ tenure + MonthlyCharges + InternetService + PaymentMethod +
          Contract + OnlineSecurity + TechSupport + PaperlessBilling,
  data = train.resamp,
  method = "knn",
  tuneLength = 10,      # Automatically tune k from a grid of 10 values
  metric = "ROC",       # Optimize for ROC metric
  trControl = ctrl
)
knn.preds <- predict(knn.fit, telecom.test)
knn.cm <- data.frame(knn=confusionMatrix(knn.preds, telecom.test$Churn,
                                         positive = "Yes", mode = "everything")$byClass)
confusionMatrix(knn.preds, telecom.test$Churn, positive = "Yes", mode = "everything")
```
````

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1095  125
       Yes  453  435

               Accuracy : 0.7258
                 95% CI : (0.7062, 0.7448)
    No Information Rate : 0.7343
    P-Value [Acc > NIR] : 0.8194

                  Kappa : 0.4079

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7768
            Specificity : 0.7074
         Pos Pred Value : 0.4899
         Neg Pred Value : 0.8975
              Precision : 0.4899
                 Recall : 0.7768
                     F1 : 0.6008
             Prevalence : 0.2657
         Detection Rate : 0.2064
   Detection Prevalence : 0.4213
      Balanced Accuracy : 0.7421

       'Positive' Class : Yes
```

The K-Nearest Neighbors (KNN) classifier achieves an accuracy of **72.5%**, which is slightly below the Random Forest and logistic regression models. The precision of the model is 48%, indicating that nearly half of the positive predictions are correct. The recall is 77%, which is lower than the the Random Forest model and logistic regression model.

CHAPTER VI

MODEL EVALUATION AND ROC CURVES

TELECOM CHURN ANALYSIS

6.1 Model Evaluation

Model Performance on the Test Set

```{r}
res.cm <- data.frame(glm.cm, svm.cm, rf.cm, knn.cm) %>%
  rename("Random Forest" = rf)
res <- data.frame(t(res.cm))
rownames(res) <- colnames(res.cm)
colnames(res) <- rownames(res.cm)
res[,c(7,5,6,2,11)] %>%
  arrange(desc(F1)) %>%
  mutate_all(percent_format(accuracy = 0.1))
```

	F1	Precision	Recall	Specificity	Balanced Accuracy
Logistic	64.3%	53.4%	80.7%	74.5%	77.6%
Random Forest	61.1%	50.2%	78.0%	72.0%	75.0%
knn	60.1%	49.0%	77.7%	70.7%	74.2%
SVM	60.1%	46.4%	85.0%	64.5%	74.8%

Out of the four models, logistic regression produces the highest F1 score, which represents the balance between precision and recall, as well as the highest specificity, which measures how well the model identifies negative cases correctly.

6.2 ROC Curves

As a final step in model selection, I will plot the ROC curves of each model with their corresponding Area Under the Curve (AUC). The Area Under the Curve measures the model's performance across all possible classification thresholds. A higher AUC indicates the model is better able to distinguish between the classes.

```{r}
Logistic <- predict(glm.fit, telecom.test, type = "prob")[,2]

SVM <- predict(svm.fit, telecom.test, type = "prob")[,2]

RandomForest <- predict(rf.fit, telecom.test, type = "prob")[,2]

KNN <- predict(knn.fit, telecom.test, type = "prob")[,2]


roc.data <- cbind(telecom.test[,20], Logistic, SVM, RandomForest, KNN)
```

ROC Curve Comparison on the Test Set

```{r}
# Ensure "Churn" is the target variable in the test set

roc.data <- data.frame(Churn = telecom.test$Churn, Logistic, SVM, RandomForest, KNN)


# Reshape the data for ROC plotting using tidyr

library(tidyr)

library(dplyr)


# Reshape the data into a long format

roc.long <- roc.data %>%
```

```
  pivot_longer(cols = c(Logistic, SVM, RandomForest, KNN),
               names_to = "Model",
               values_to = "Prediction")

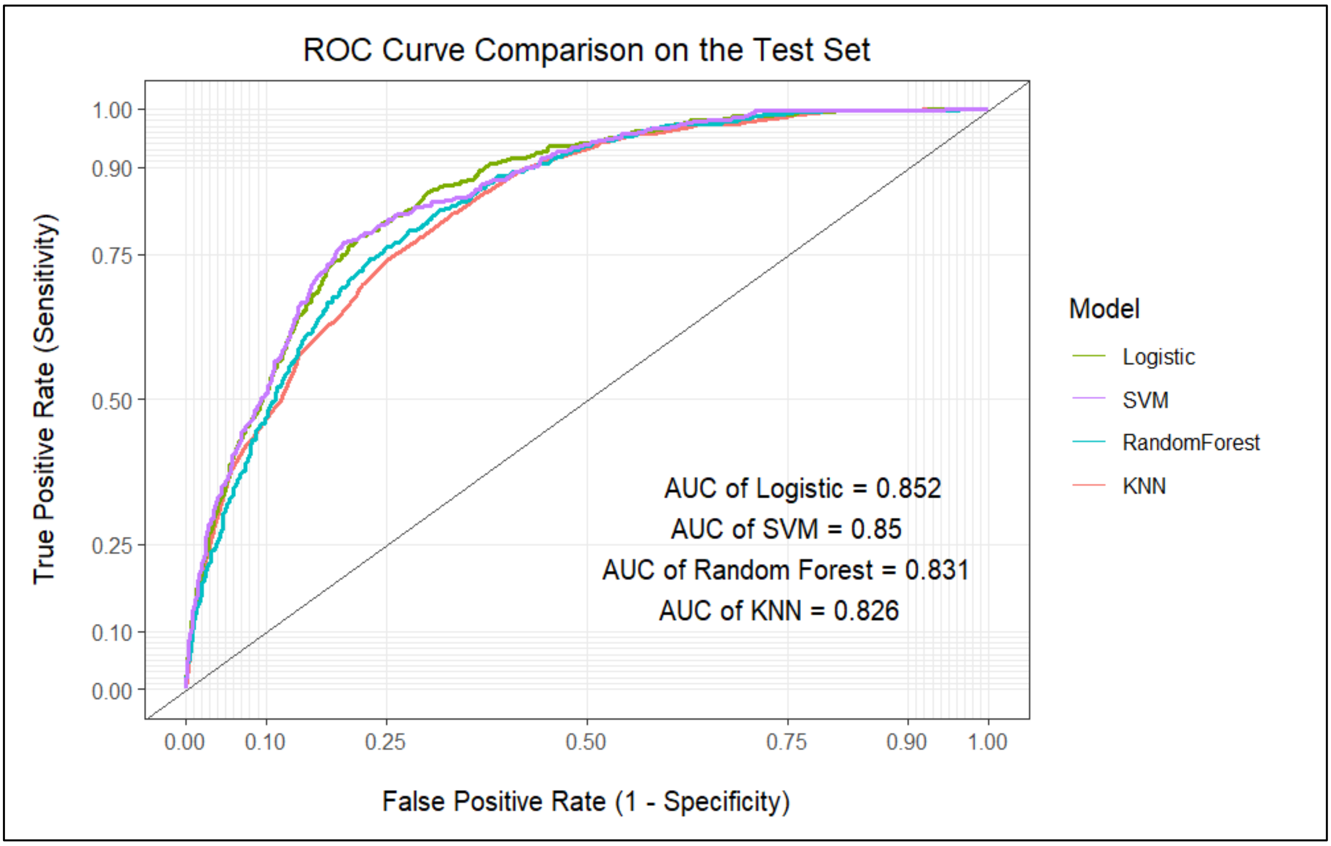
# Convert Churn to binary values (1 for "Yes", 0 for "No")
roc.long$Churn <- ifelse(roc.long$Churn == "Yes", 1, 0)

# Check the structure of the reshaped data
str(roc.long)

# Now generate the ROC plot
rocplot <- ggplot(roc.long, aes(d = Churn, m = Prediction, color = Model)) +
  geom_roc(n.cuts = 0) +
  style_roc(xlab = "\nFalse Positive Rate (1 - Specificity)",
            ylab = "True Positive Rate (Sensitivity)\n") +
  labs(title = "ROC Curve Comparison on the Test Set", color = "Model") +
  theme(plot.title = element_text(hjust = 0.5))

# Add AUC values and abline
rocplot +
  geom_abline(size = 0.5, color = "grey30") +
  annotate("text", x = 0.77, y = 0.35, label = paste("AUC of Logistic =", round(calc_auc(rocplot)$AUC[2], 3))) +
```

```
 annotate("text", x = 0.75, y = 0.28, label = paste("AUC of SVM =", round(calc_auc(rocplot)$AUC[4], 3))) +
 annotate("text", x = 0.75, y = 0.21, label = paste("AUC of Random Forest =", round(calc_auc(rocplot)$AUC[3], 3))) +
 annotate("text", x = 0.74, y = 0.14, label = paste("AUC of KNN =", round(calc_auc(rocplot)$AUC[1], 3))) +
 scale_color_discrete(breaks = c("Logistic", "SVM", "RandomForest", "KNN"))
```



Out of the four classifiers, the logistic model has the highest Area Under the Curve of 0.854 on the test set. This represents the probability that our model will rate or rank a randomly chosen observation from the positive class, Churn = Yes, as more likely to be from that class than a randomly chosen nonpositive observation, Churn = No (Hanley & McNeil, 1982).

6.3 Key Findings

Overall, the logistic regression model had the strongest performance on the test set. Based on the coefficients from the model, at least one category in all eight predictors has a significant association to customer attrition. A summary of the relationships of each, when all other variables are held constant, is listed in the table below.

```{r}
glm.fit <- train(Churn ~ tenure + MonthlyCharges + InternetService + PaymentMethod +
                 Contract + OnlineSecurity + TechSupport + PaperlessBilling,
                 data = telco, method = "glm",
                 preProcess = c("center", "scale"),
                 trControl = trainControl(method = "cv", number = 10))
```

```
Call:
NULL

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                           -1.61343    0.04656 -34.656  < 2e-16 ***
tenure                                -0.76083    0.05258 -14.470  < 2e-16 ***
MonthlyCharges                         0.32805    0.09860   3.327 0.000878 ***
`InternetServiceFiber optic`           0.26989    0.06954   3.881 0.000104 ***
InternetServiceNo                     -0.32844    0.06323  -5.195 2.05e-07 ***
`PaymentMethodCredit card (automatic)` -0.03354   0.04659  -0.720 0.471601
`PaymentMethodElectronic check`        0.17718    0.04417   4.011 6.04e-05 ***
`PaymentMethodMailed check`           -0.01368    0.04728  -0.289 0.772271
`ContractOne year`                    -0.28241    0.04279  -6.600 4.13e-11 ***
`ContractTwo year`                    -0.58727    0.07400  -7.936 2.09e-15 ***
OnlineSecurityYes                     -0.21173    0.03854  -5.494 3.93e-08 ***
TechSupportYes                        -0.18553    0.03982  -4.659 3.18e-06 ***
PaperlessBillingYes                    0.18983    0.03610   5.259 1.45e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8143.4  on 7031  degrees of freedom
Residual deviance: 5922.9  on 7019  degrees of freedom
AIC: 5948.9

Number of Fisher Scoring iterations: 6
```

```
summary(glm.fit$finalModel)
```

Extracting Coefficients and Calculating Odds Ratios

````
```{r}
OR <- coef(glm.fit$finalModel) %>% exp() %>% round(digits = 2) %>%
as.data.frame() %>% slice(-c(1,6,8))
data.frame(Predictor = c("Tenure", "MonthlyCharges", "InternetServiceFibreOptic",
                     "InternetServiceNo", "PaymentMethodECheck", "ContractOneYear",
"ContractTwoYear",
                     "OnlineSecurityYes", "TechSupportYes", "PaperlessBillingYes"),
        OddsRatio = OR[,1],
        Interpretation = c("A one month increase in tenure decreases the risk of
churning by about 53%.",
                          "For every $1 increase in monthly charges, we expect to see an
increase in
                          the odds of churning by a factor of 1.39 or by 39%.",
                          "Customers with fibre optic internet are 31% more likely to churn
than those
                          with DSL.", "Those without internet are 28% less likely to churn
than
                          customers with DSL internet.", "Customers who pay with
electronic checks are
                          more likely to churn by a factor of 1.19 or by 19% compared to
customers who use
                          automatic bank transfers.", "Customers on one-year contracts are
25% less likely
                          to churn than customers on month-to-month contracts.",
"Customers
                          on two-year contracts are 44% less likely to churn compared to
those on
                          month-to-month contracts.", "Customers with online security are
19% less likely
                          to churn than customers without online security.", "Customers
with tech support
                          are about 17% less likely to churn than customers without tech
support.",
````
````

````
                              "Customers with paperless billing are 21% more likely to churn
than customers
                              without paperless billing.")) %>%
  arrange(desc(OddsRatio)) %>% view()

```
````

	Predictor	OddsRatio	Interpretation
1	MonthlyCharges	1.39	For every $1 increase in monthly charges, we expect to see ...
2	InternetServiceFiberOptic	1.31	Customers with fiber optic internet are 31% more likely to c...
3	PaperlessBillingYes	1.21	Customers with paperless billing are 21% more likely to chu...
4	PaymentMethodECheck	1.19	Customers who pay with electronic checks are ...
5	TechSupportYes	0.83	Customers with tech support are about 1...
6	OnlineSecurityYes	0.81	Customers with online security are 19% less likely ...
7	ContractOneYear	0.75	Customers on one-year contracts are 25% less likely ...
8	InternetServiceNo	0.72	Those without internet are 28% less likely to churn than ...
9	ContractTwoYear	0.56	Customers on two-year contracts are 44...
10	Tenure	0.47	A one month increase in tenure decreases the risk of churni...
````

CHAPTER VII
PREDICTION AND INFERENCE
TELECOM CHURN ANALYSIS

7.1 Prediction

Based on the model output, the following key predictions have been made regarding customer churn:

- The model achieved an accuracy of **76.14%**, indicating a reasonable level of reliability in predicting customer churn.
- Sensitivity (recall for churned customers) is **80.71%**, meaning the model effectively identifies a significant portion of customers likely to churn.
- Specificity (ability to identify non-churned customers) stands at **74.48%**, ensuring balanced prediction performance.
- Precision for predicting churn is **53.36%**, signifying that 53.36% of customers predicted to churn actually do so.
- The F1-score of **64.25%** reflects a balance between precision and recall, ensuring a robust prediction model.

These insights suggest that the model can be used effectively for churn prevention strategies, identifying high-risk customers and enabling targeted retention efforts.

7.2 Inference

From the analysis and model results, the following inferences can be drawn:

- **Service Usage Impact:** Customers with fewer subscribed services tend to have higher churn rates, as shown by the declining churn rate with increasing services.
- **Tenure Effect:** Customers with lower tenure are more likely to churn, whereas long-tenure customers show greater retention.
- **Service Combination Influence:** Customers who use both phone and internet services exhibit a higher churn rate than those using only one.

- **Internet Service Type:** Fibre optic users have the highest churn rate compared to DSL and customers without internet service.
- **Model Reliability:** While the model is effective in identifying potential churners, its precision indicates room for improvement, possibly through additional feature engineering or hyper parameter tuning.

These insights can guide strategic interventions, such as personalized retention offers, improved service bundles, and proactive engagement to reduce customer churn.

CHAPTER VII
CONCLUSION
TELECOM CHURN ANALYSIS

7.1 Conclusion

In predicting customer attrition, logistic regression produced the highest Area Under the Curve, F1 score, and specificity. Some of the most important predictors of customer attrition include Tenure, MonthlyCharges, InternetService, PaymentMethod, Contract, OnlineSecurity, TechSupport, and PaperlessBilling. We also found that the most significant relationships from our logistic model are the customer’s monthly charges, the type of internet service and contract they have, and the length of time they have been customers with Telco. To proactively reduce their churn rate, Telco could target customers who are on month-to-month contracts, use fibre optic internet, have higher monthly charges on average, and who have a shorter tenure of less than 18 months, which is the average tenure of their former customers.

REFERENCES

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). *Smote: Synthetic minority over-sampling technique*. Journal of Artificial Intelligence Research, 16:321-357.

Hanley, J. A., & Mcneil, B. J. (1982). *The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve*. Radiology, 143(1), 29-36. doi:10.1148/radiology.143.1.7063747

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York, NY: Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. New York, NY: Springer.

Torgo, L. (2010) *Data Mining using R: learning with case studies*, CRC Press (ISBN: 9781439810187). http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR