

# Project

Vishwas Ramakrishna (A20552892)      Vrushabh Shet (A20560742)  
Rachana Vijay (A20605843)      Sathvika Reddy Madithati (A20584937)  
Umesh Chand Vankayalapati (A20576496)

12-05-2025

```
# Core Libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr     1.1.4      v readr     2.1.5
## vforcats   1.0.0      v stringr   1.5.1
## v ggplot2   4.0.0      v tibble    3.3.0
## v lubridate 1.9.4      v tidyr    1.3.1
## v purrr    1.1.0
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()   masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
##
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
##
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
##
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
# Visualization
library(ggplot2)
library(cowplot)
```

```
##
```

```
## Attaching package: 'cowplot'  
##  
## The following object is masked from 'package:lubridate':  
##  
##     stamp
```

```
library(gridExtra)
```

```
##  
## Attaching package: 'gridExtra'  
##  
## The following object is masked from 'package:dplyr':  
##  
##     combine
```

```
# Data Preprocessing  
library(caret)
```

```
## Loading required package: lattice  
##  
## Attaching package: 'caret'  
##  
## The following object is masked from 'package:purrr':  
##  
##     lift
```

```
library(recipes)
```

```
##  
## Attaching package: 'recipes'  
##  
## The following object is masked from 'package:stringr':  
##  
##     fixed  
##  
## The following object is masked from 'package:stats':  
##  
##     step
```

```
library(dplyr)  
library(forcats)
```

```
# Unsupervised Learning  
library(stats)
```

```
# Modeling and Evaluation  
library(caret)  
library(randomForest)
```

```
## randomForest 4.7-1.2  
## Type rfNews() to see new features/changes/bug fixes.
```

```
##  
## Attaching package: 'randomForest'  
##  
## The following object is masked from 'package:gridExtra':  
##  
##     combine  
##  
## The following object is masked from 'package:dplyr':  
##  
##     combine  
##  
## The following object is masked from 'package:ggplot2':  
##  
##     margin
```

```
library(gbm)
```

```
## Loaded gbm 2.2.2  
## This version of gbm is no longer under development. Consider transitioning to gbm3, https://github.com
```

```
library(e1071)
```

```
##  
## Attaching package: 'e1071'  
##  
## The following object is masked from 'package:ggplot2':  
##  
##     element
```

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.  
##  
## Attaching package: 'pROC'  
##  
## The following objects are masked from 'package:stats':  
##  
##     cov, smooth, var
```

```
library(MASS)
```

```
##  
## Attaching package: 'MASS'  
##  
## The following object is masked from 'package:dplyr':  
##  
##     select
```

```
library(cluster)  
library(FactoMineR)  
library(factoextra)
```

```

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(conflicted)

# Load required libraries
library(tidyverse)
library(data.table)

# Load the dataset
file_path <- "/Users/vishwasramakrishna/Downloads/DPA_Project_Telco_Customer_Churn_Prediction 2/telco.csv"
telco_data <- fread(file_path)

# View dimensions of the dataset
dim(telco_data)

## [1] 7043   50

# Check for missing values
missing_summary <- colSums(is.na(telco_data))
cat("Missing Values per Column:\n")

## Missing Values per Column:

print(missing_summary)

##          Customer ID           Gender
##                         0
##                      Age Under 30
##                         0
## Senior Citizen       Married
##                         0
## Dependents      Number of Dependents
##                         0
##                      Country           State
##                         0
##                      City        Zip Code
##                         0
##                     Latitude      Longitude
##                         0
## Population        Quarter
##                         0
## Referred a Friend Number of Referrals
##                         0
## Tenure in Months           Offer
##                         0
## Phone Service Avg Monthly Long Distance Charges
##                         0
## Multiple Lines           Internet Service
##                         0
## Internet Type Avg Monthly GB Download
##                         0

```

```

##          Online Security          Online Backup
##                      0                      0
##          Device Protection Plan      Premium Tech Support
##                      0                      0
##          Streaming TV              Streaming Movies
##                      0                      0
##          Streaming Music           Unlimited Data
##                      0                      0
##          Contract                  Paperless Billing
##                      0                      0
##          Payment Method             Monthly Charge
##                      0                      0
##          Total Charges              Total Refunds
##                      0                      0
##          Total Extra Data Charges   Total Long Distance Charges
##                      0                      0
##          Total Revenue              Satisfaction Score
##                      0                      0
##          Customer Status            Churn Label
##                      0                      0
##          Churn Score                CLTV
##                      0                      0
##          Churn Category             Churn Reason
##                      0                      0

```

```

# Check percentage of missing values
missing_percentage <- (colSums(is.na(telco_data)) / nrow(telco_data)) * 100
cat("\nPercentage of Missing Values:\n")

```

```

## Percentage of Missing Values:

```

```
print(missing_percentage)
```

##	Customer ID	Gender
##	0	0
##	Age	Under 30
##	0	0
##	Senior Citizen	Married
##	0	0
##	Dependents	Number of Dependents
##	0	0
##	Country	State
##	0	0
##	City	Zip Code
##	0	0
##	Latitude	Longitude
##	0	0
##	Population	Quarter
##	0	0
##	Referred a Friend	Number of Referrals
##	0	0
##	Tenure in Months	Offer

```

##          0          0
##      Phone Service Avg Monthly Long Distance Charges
##          0          0
##      Multiple Lines           Internet Service
##          0          0
##      Internet Type Avg Monthly GB Download
##          0          0
##      Online Security           Online Backup
##          0          0
##      Device Protection Plan Premium Tech Support
##          0          0
##      Streaming TV           Streaming Movies
##          0          0
##      Streaming Music           Unlimited Data
##          0          0
##      Contract           Paperless Billing
##          0          0
##      Payment Method           Monthly Charge
##          0          0
##      Total Charges           Total Refunds
##          0          0
##      Total Extra Data Charges Total Long Distance Charges
##          0          0
##      Total Revenue           Satisfaction Score
##          0          0
##      Customer Status           Churn Label
##          0          0
##      Churn Score           CLTV
##          0          0
##      Churn Category           Churn Reason
##          0          0

```

```

# Handle missing values in 'Offer' column
telco_data$Offer[is.na(telco_data$Offer)] <- "No Offer"

```

```

# Verify missing values for 'Offer' after handling
missing_offer <- sum(is.na(telco_data$Offer))
cat("Missing Values in 'Offer' After Handling:", missing_offer, "\n")

```

```

## Missing Values in 'Offer' After Handling: 0

```

```

# Handle missing values in 'Internet Type' using the most frequent value (mode)
most_frequent_internet_type <- names(sort(table(telco_data$`Internet Type`), decreasing = TRUE))[1]
telco_data$`Internet Type`[is.na(telco_data$`Internet Type`)] <- most_frequent_internet_type

```

```

# Verify missing values for 'Internet Type' after handling
missing_internet_type <- sum(is.na(telco_data$`Internet Type`))
cat("Missing Values in 'Internet Type' After Handling:", missing_internet_type, "\n")

```

```

## Missing Values in 'Internet Type' After Handling: 0

```

```

library(dplyr)

# Summary statistics for numeric columns
summary(as.data.frame(telco_data)[sapply(telco_data, is.numeric)])

```

	Age	Number of Dependents	Zip Code	Latitude
##	Min. :19.00	Min. :0.0000	Min. :90001	Min. :32.56
##	1st Qu.:32.00	1st Qu.:0.0000	1st Qu.:92101	1st Qu.:33.99
##	Median :46.00	Median :0.0000	Median :93518	Median :36.21
##	Mean :46.51	Mean :0.4687	Mean :93486	Mean :36.20
##	3rd Qu.:60.00	3rd Qu.:0.0000	3rd Qu.:95329	3rd Qu.:38.16
##	Max. :80.00	Max. :9.0000	Max. :96150	Max. :41.96
##	Longitude	Population	Number of Referrals	Tenure in Months
##	Min. :-124.3	Min. : 11	Min. : 0.000	Min. : 1.00
##	1st Qu.:-121.8	1st Qu.: 2344	1st Qu.: 0.000	1st Qu.: 9.00
##	Median :-119.6	Median : 17554	Median : 0.000	Median :29.00
##	Mean :-119.8	Mean : 22140	Mean : 1.952	Mean :32.39
##	3rd Qu.:-118.0	3rd Qu.: 36125	3rd Qu.: 3.000	3rd Qu.:55.00
##	Max. :-114.2	Max. :105285	Max. :11.000	Max. :72.00
##	Avg Monthly Long Distance Charges	Avg Monthly GB Download	Monthly Charge	
##	Min. : 0.00	Min. : 0.00	Min. : 18.25	
##	1st Qu.: 9.21	1st Qu.: 3.00	1st Qu.: 35.50	
##	Median :22.89	Median :17.00	Median : 70.35	
##	Mean :22.96	Mean :20.52	Mean : 64.76	
##	3rd Qu.:36.40	3rd Qu.:27.00	3rd Qu.:89.85	
##	Max. :49.99	Max. :85.00	Max. :118.75	
##	Total Charges	Total Refunds	Total Extra Data Charges	
##	Min. : 18.8	Min. : 0.000	Min. : 0.000	
##	1st Qu.: 400.1	1st Qu.: 0.000	1st Qu.: 0.000	
##	Median :1394.5	Median : 0.000	Median : 0.000	
##	Mean :2280.4	Mean : 1.962	Mean : 6.861	
##	3rd Qu.:3786.6	3rd Qu.: 0.000	3rd Qu.: 0.000	
##	Max. :8684.8	Max. :49.790	Max. :150.000	
##	Total Long Distance Charges	Total Revenue	Satisfaction Score	
##	Min. : 0.00	Min. : 21.36	Min. :1.000	
##	1st Qu.: 70.55	1st Qu.: 605.61	1st Qu.:3.000	
##	Median : 401.44	Median : 2108.64	Median :3.000	
##	Mean : 749.10	Mean : 3034.38	Mean :3.245	
##	3rd Qu.:1191.10	3rd Qu.: 4801.15	3rd Qu.:4.000	
##	Max. :3564.72	Max. :11979.34	Max. :5.000	
##	Churn Score	CLTV		
##	Min. : 5.00	Min. :2003		
##	1st Qu.:40.00	1st Qu.:3469		
##	Median :61.00	Median :4527		
##	Mean :58.51	Mean :4400		
##	3rd Qu.:75.50	3rd Qu.:5380		
##	Max. :96.00	Max. :6500		

```

# Check for missing values per column
missing_summary <- colSums(is.na(telco_data))
cat("Missing Values per Column:\n")

```

## Missing Values per Column:

```
print(missing_summary)
```

```
## Customer ID Gender
## 0 0
## Age Under 30
## 0 0
## Senior Citizen Married
## 0 0
## Dependents Number of Dependents
## 0 0
## Country State
## 0 0
## City Zip Code
## 0 0
## Latitude Longitude
## 0 0
## Population Quarter
## 0 0
## Referred a Friend Number of Referrals
## 0 0
## Tenure in Months Offer
## 0 0
## Phone Service Avg Monthly Long Distance Charges
## 0 0
## Multiple Lines Internet Service
## 0 0
## Internet Type Avg Monthly GB Download
## 0 0
## Online Security Online Backup
## 0 0
## Device Protection Plan Premium Tech Support
## 0 0
## Streaming TV Streaming Movies
## 0 0
## Streaming Music Unlimited Data
## 0 0
## Contract Paperless Billing
## 0 0
## Payment Method Monthly Charge
## 0 0
## Total Charges Total Refunds
## 0 0
## Total Extra Data Charges Total Long Distance Charges
## 0 0
## Total Revenue Satisfaction Score
## 0 0
## Customer Status Churn Label
## 0 0
## Churn Score CLTV
## 0 0
## Churn Category Churn Reason
## 0 0
```

```

# Handle missing values for 'Churn Category'
telco_data$`Churn Category`[is.na(telco_data$`Churn Category`)] <- "Not Applicable"

# Verify missing values for 'Churn Category'
missing_churn_category <- sum(is.na(telco_data$`Churn Category`))
cat("Missing Values in 'Churn Category' After Handling:", missing_churn_category, "\n")

## Missing Values in 'Churn Category' After Handling: 0

# Handle missing values for 'Churn Reason'
telco_data$`Churn Reason`[is.na(telco_data$`Churn Reason`)] <- "Not Applicable"

# Verify missing values for 'Churn Reason'
missing_churn_reason <- sum(is.na(telco_data$`Churn Reason`))
cat("Missing Values in 'Churn Reason' After Handling:", missing_churn_reason, "\n")

## Missing Values in 'Churn Reason' After Handling: 0

# Final check for missing values per column
missing_summary <- colSums(is.na(telco_data))
cat("Missing Values per Column:\n")

## Missing Values per Column:

print(missing_summary)

##          Customer ID           Gender
##                         0                         0
##                      Age      Under 30
##                         0                         0
##        Senior Citizen       Married
##                         0                         0
##        Dependents   Number of Dependents
##                         0                         0
##          Country            State
##                         0                         0
##          City              Zip Code
##                         0                         0
##        Latitude      Longitude
##                         0                         0
##        Population        Quarter
##                         0                         0
##    Referred a Friend  Number of Referrals
##                         0                         0
##        Tenure in Months          Offer
##                         0                         0
##          Phone Service Avg Monthly Long Distance Charges
##                         0                         0
##        Multiple Lines       Internet Service
##                         0                         0
##        Internet Type Avg Monthly GB Download

```

```

##          0          0
##          Online Security      Online Backup
##          0          0
##          Device Protection Plan      Premium Tech Support
##          0          0
##          Streaming TV      Streaming Movies
##          0          0
##          Streaming Music      Unlimited Data
##          0          0
##          Contract      Paperless Billing
##          0          0
##          Payment Method      Monthly Charge
##          0          0
##          Total Charges      Total Refunds
##          0          0
##          Total Extra Data Charges      Total Long Distance Charges
##          0          0
##          Total Revenue      Satisfaction Score
##          0          0
##          Customer Status      Churn Label
##          0          0
##          Churn Score      CLTV
##          0          0
##          Churn Category      Churn Reason
##          0          0

```

```

# Visualization of the target variable
churn_counts <- table(telco_data$`Churn Label`)
print(churn_counts)

```

```

##
##    No   Yes
## 5174 1869

```

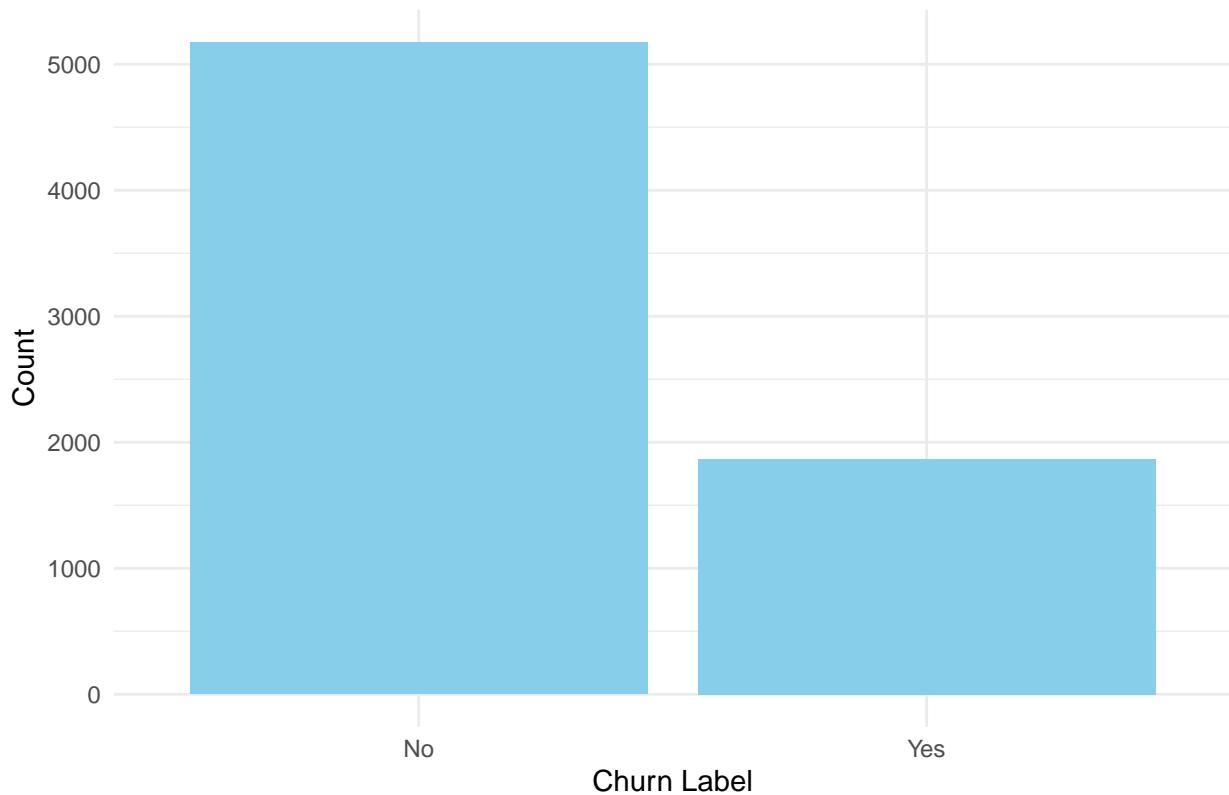
```

# Load ggplot2 for visualization
library(ggplot2)

# Bar plot of 'Churn Label'
ggplot(telco_data, aes(x = `Churn Label`)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Churn Label Distribution", x = "Churn Label", y = "Count") +
  theme_minimal()

```

## Churn Label Distribution



```
# Load necessary libraries
library(dplyr)
library(ggcorrplot)
library(ggplot2)

telco_df <- as.data.frame(telco_data)

numeric_data <- select_if(telco_df, is.numeric)

# Compute correlation matrix
correlation_matrix <- cor(numeric_data, use = "complete.obs")

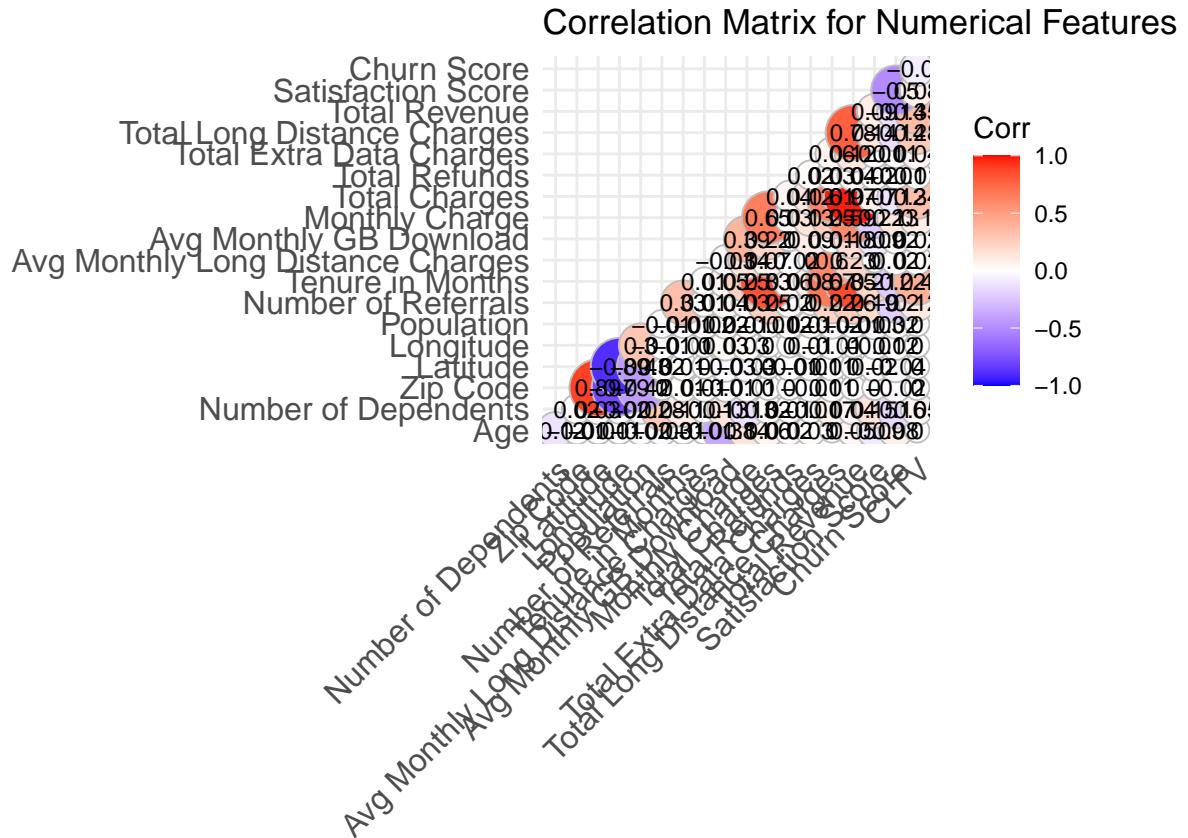
# Plot heatmap
ggcorrplot(correlation_matrix,
            method = "circle",
            type = "lower",
            lab = TRUE,
            lab_size = 3,
            colors = c("blue", "white", "red"),
            title = "Correlation Matrix for Numerical Features",
            ggtheme = theme_minimal())

## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()``.
## i See also `vignette("ggplot2-in-packages")` for more information.
```

```

## i The deprecated feature was likely used in the ggcormplot package.
## Please report the issue at <https://github.com/kassambara/ggcormplot/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

```



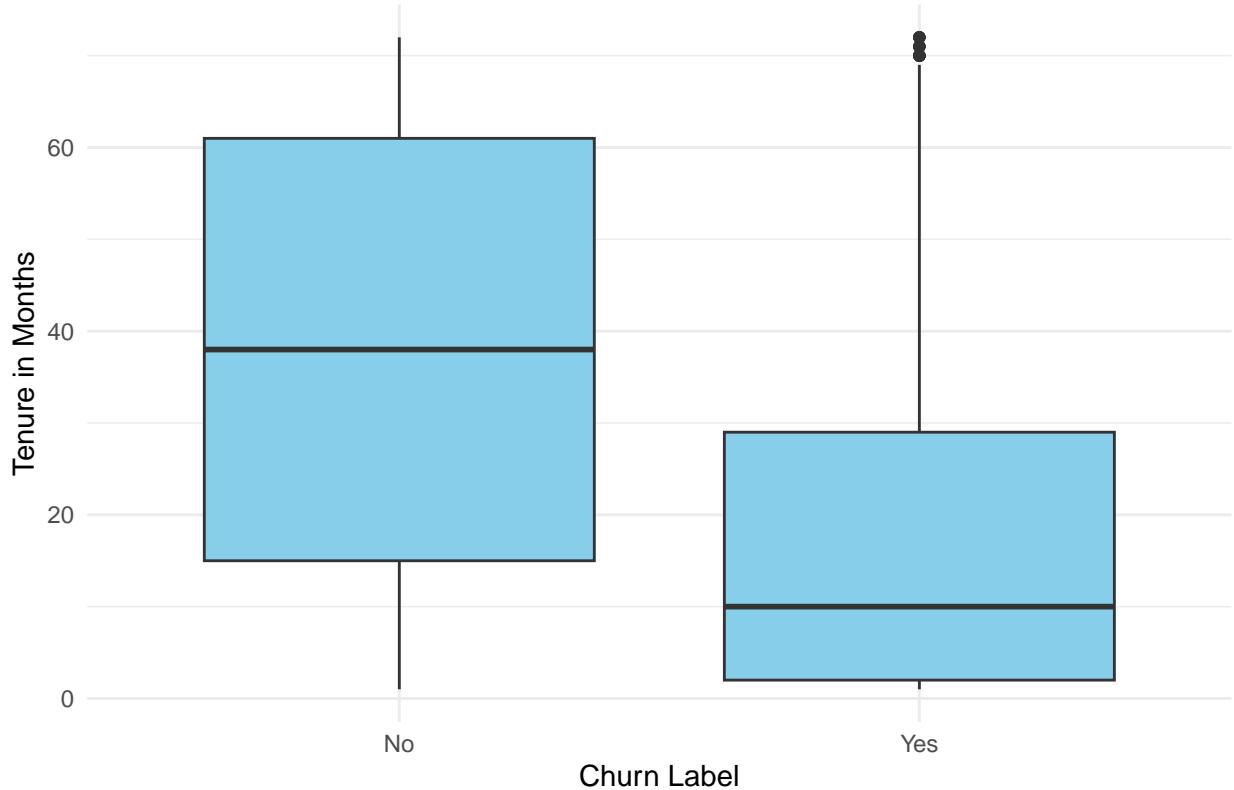
```

library(ggplot2)

# Boxplot: Tenure in Months vs Churn Label
ggplot(telco_data, aes(x = `Churn Label`, y = `Tenure in Months`)) +
  geom_boxplot(fill = "skyblue") +
  labs(title = "Tenure in Months by Churn Label", x = "Churn Label", y = "Tenure in Months") +
  theme_minimal()

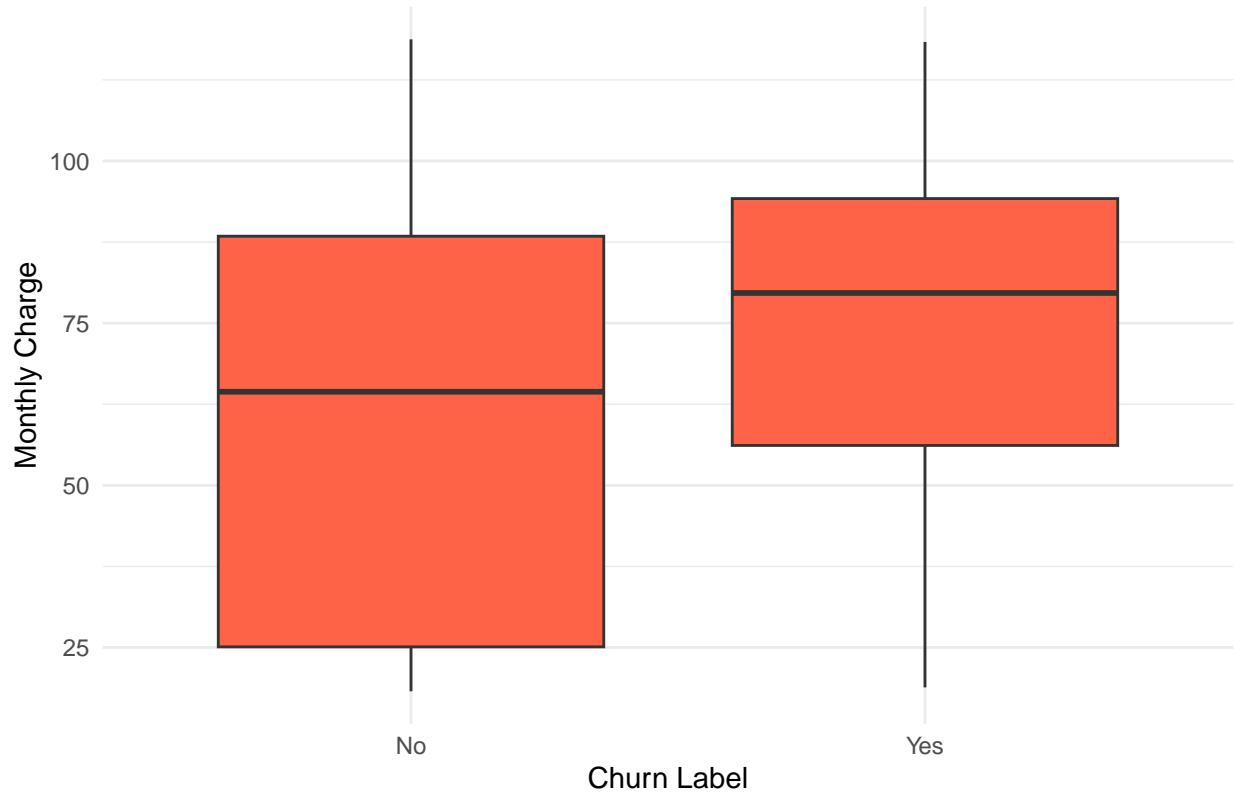
```

### Tenure in Months by Churn Label



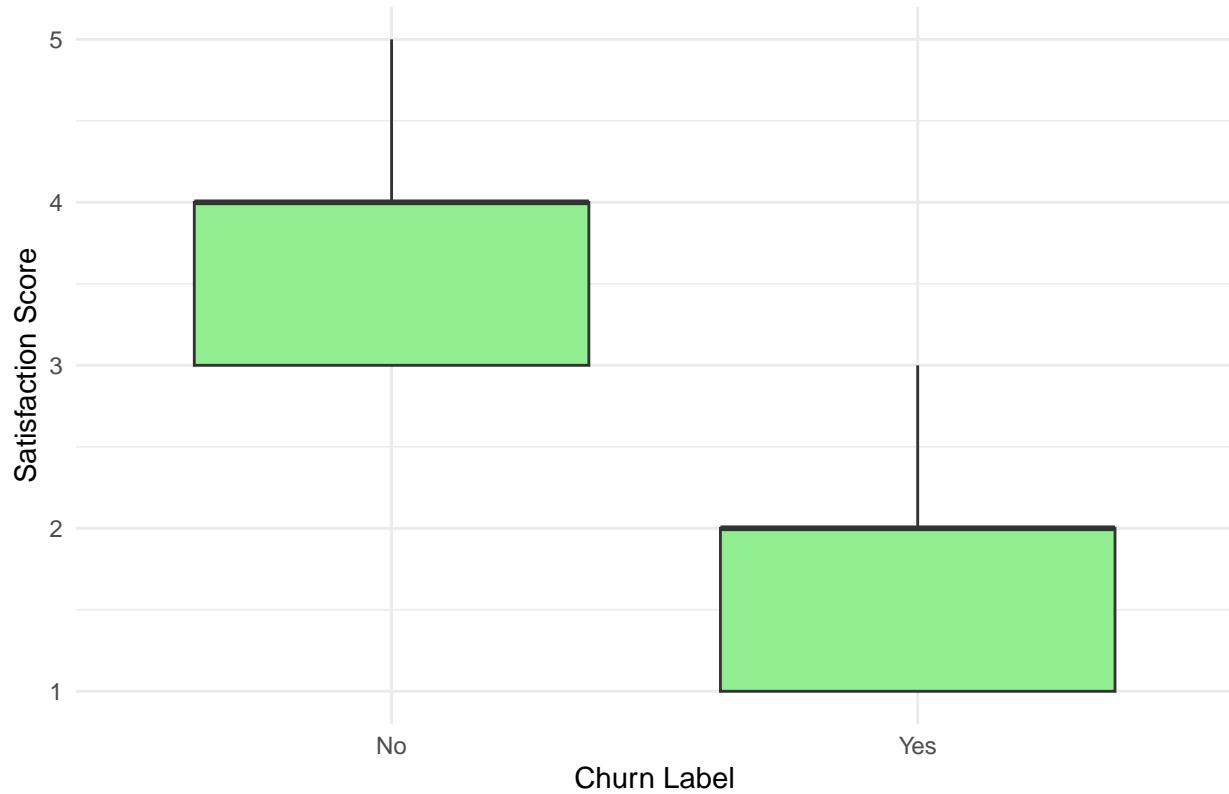
```
# Boxplot: Monthly Charges vs Churn Label
ggplot(telco_data, aes(x = `Churn Label`, y = `Monthly Charge`)) +
  geom_boxplot(fill = "tomato") +
  labs(title = "Monthly Charges by Churn Label", x = "Churn Label", y = "Monthly Charge") +
  theme_minimal()
```

## Monthly Charges by Churn Label



```
# Boxplot: Satisfaction Score vs Churn Label
ggplot(telco_data, aes(x = `Churn Label`, y = `Satisfaction Score`)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Satisfaction Score by Churn Label", x = "Churn Label", y = "Satisfaction Score") +
  theme_minimal()
```

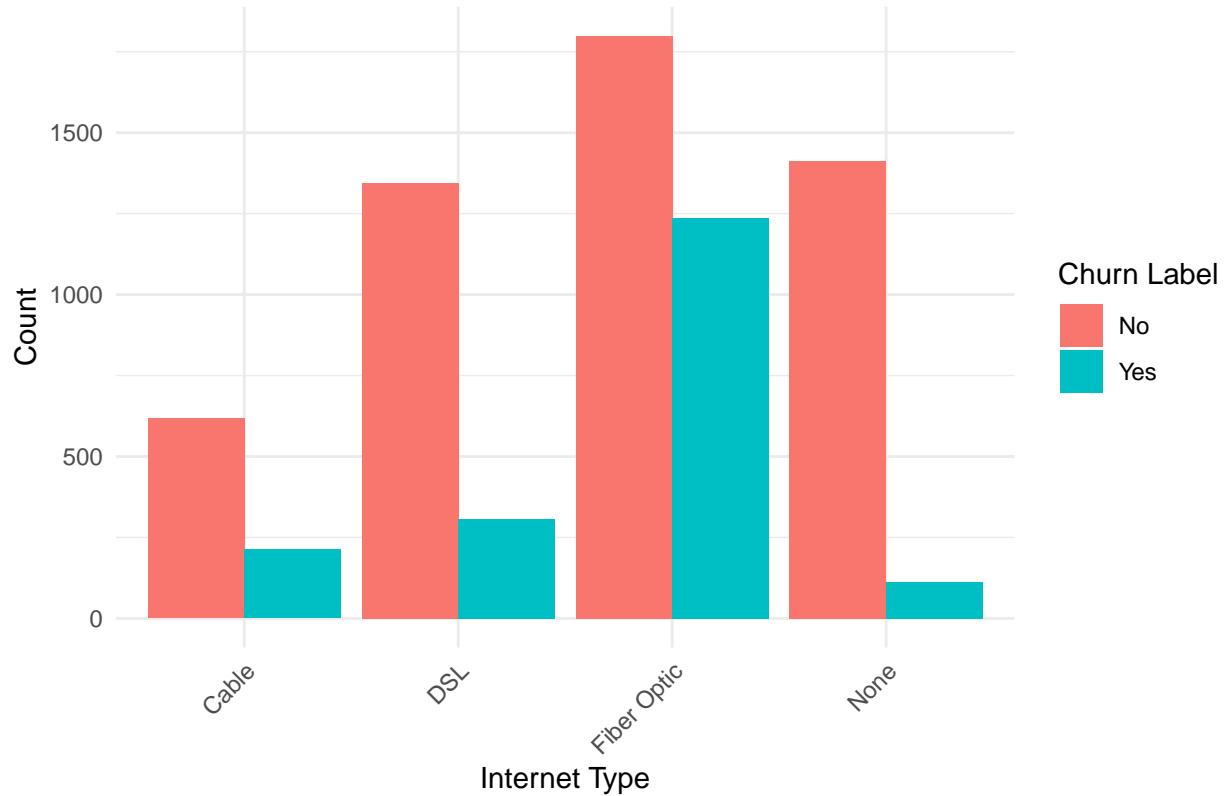
## Satisfaction Score by Churn Label



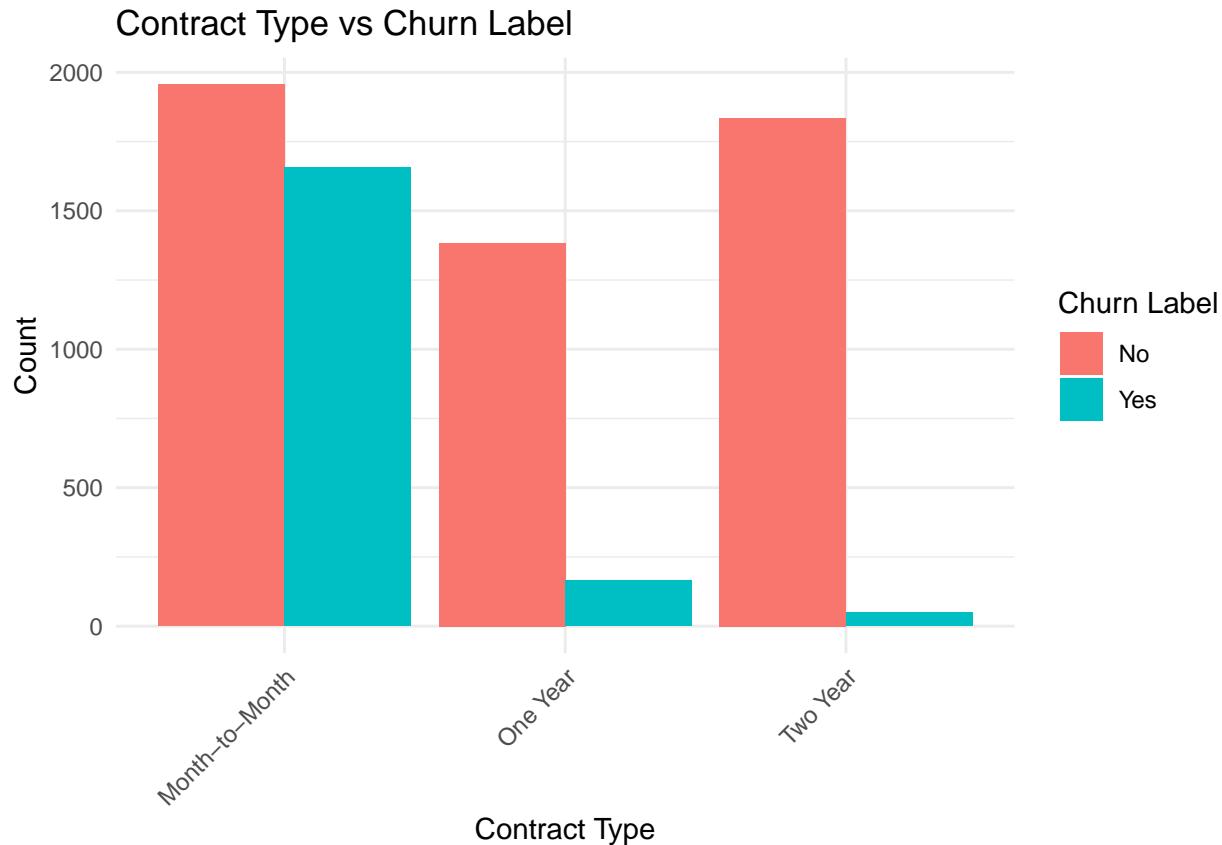
```
library(ggplot2)

# Countplot: Internet Type vs Churn Label
ggplot(telco_data, aes(x = `Internet Type`, fill = `Churn Label`)) +
  geom_bar(position = "dodge") +
  labs(title = "Internet Type vs Churn Label", x = "Internet Type", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

### Internet Type vs Churn Label



```
# Countplot: Contract Type vs Churn Label
ggplot(telco_data, aes(x = Contract, fill = `Churn Label`)) +
  geom_bar(position = "dodge") +
  labs(title = "Contract Type vs Churn Label", x = "Contract Type", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
# Chi-squared test: Internet Type vs Churn Label
internet_type_churn <- table(telco_data$`Internet Type`, telco_data$`Churn Label`)
chi_test_1 <- chisq.test(internet_type_churn)

cat("Chi-squared Test for 'Internet Type' and 'Churn Label':\n")

## Chi-squared Test for 'Internet Type' and 'Churn Label':

cat("Chi2 Stat:", chi_test_1$statistic, ", P-value:", chi_test_1$p.value,
    ", Degrees of Freedom:", chi_test_1$parameter, "\n")

## Chi2 Stat: 653.8321 , P-value: 2.15035e-141 , Degrees of Freedom: 3

if (chi_test_1$p.value < 0.05) {
  cat("There is a significant relationship between Internet Type and Churn Label.\n\n")
} else {
  cat("No significant relationship between Internet Type and Churn Label.\n\n")
}

## There is a significant relationship between Internet Type and Churn Label.

# Chi-squared test: Contract Type vs Churn Label
contract_churn <- table(telco_data$Contract, telco_data$`Churn Label`)
```

```

chi_test_2 <- chisq.test(contract_churn)

cat("Chi-squared Test for 'Contract' and 'Churn Label':\n")

## Chi-squared Test for 'Contract' and 'Churn Label':

cat("Chi2 Stat:", chi_test_2$statistic, ", P-value:", chi_test_2$p.value,
    ", Degrees of Freedom:", chi_test_2$parameter, "\n")

## Chi2 Stat: 1445.293 , P-value: 1.440655e-314 , Degrees of Freedom: 2

if (chi_test_2$p.value < 0.05) {
  cat("There is a significant relationship between Contract Type and Churn Label.\n")
} else {
  cat("No significant relationship between Contract Type and Churn Label.\n")
}

## There is a significant relationship between Contract Type and Churn Label.

library(recipes)
library(dplyr)
library(caret)

telco_df <- as.data.frame(telco_data)

# Select relevant features
selected_features <- c("Tenure in Months", "Monthly Charge", "Satisfaction Score",
                      "Contract", "Internet Type")

telco_selected <- dplyr::select(telco_df, all_of(selected_features))

# Define preprocessing recipe
rec <- recipe(~ ., data = telco_selected) %>%
  step_dummy(all_nominal_predictors(), one_hot = TRUE) %>%
  step_center(all_numeric_predictors()) %>%
  step_scale(all_numeric_predictors())

# Prep and bake (apply) the transformations
prepped_data <- prep(rec, training = telco_selected)
processed_features <- bake(prepped_data, new_data = telco_selected)

# Check number of resulting features
n_features <- ncol(processed_features)
n_components_value <- min(7, n_features)

# Output
cat("Total Features After Preprocessing:", n_features, "\n")

## Total Features After Preprocessing: 10

```

```

cat("Number of PCA Components to Use:", n_components_value, "\n")

## Number of PCA Components to Use: 7

cat("Number of Features:", ncol(processed_features), "\n")

## Number of Features: 10

# Apply PCA
pca_result <- prcomp(processed_features, center = TRUE, scale. = TRUE)

# Get the number of components
n_features <- ncol(processed_features)
n_components <- min(7, n_features)

# Compute cumulative explained variance
explained_variance <- cumsum(pca_result$sdev^2 / sum(pca_result$sdev^2))

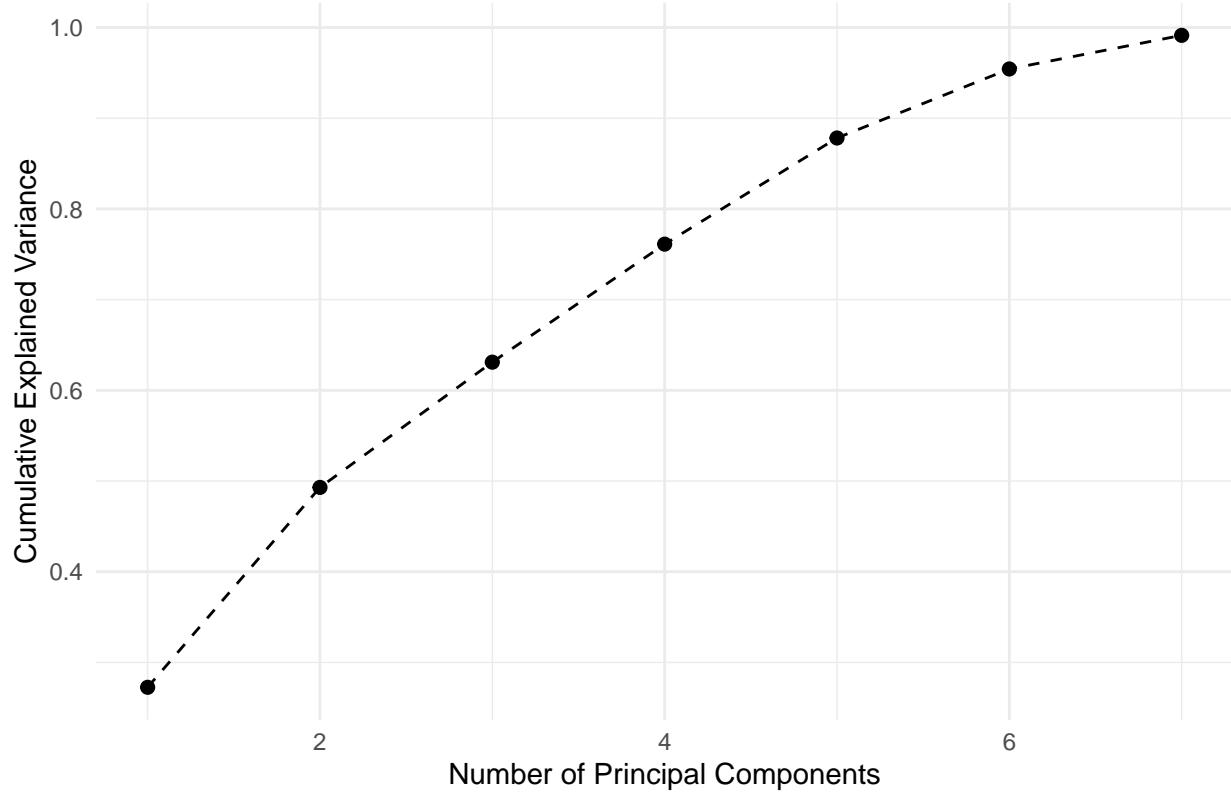
# Prepare data for plotting
pca_df <- data.frame(
  Components = 1:n_components,
  CumulativeExplainedVariance = explained_variance[1:n_components]
)

# Plot cumulative explained variance
library(ggplot2)

ggplot(pca_df, aes(x = Components, y = CumulativeExplainedVariance)) +
  geom_line(linetype = "dashed") +
  geom_point(size = 2) +
  labs(title = "Explained Variance by Principal Components",
       x = "Number of Principal Components",
       y = "Cumulative Explained Variance") +
  theme_minimal()

```

## Explained Variance by Principal Components



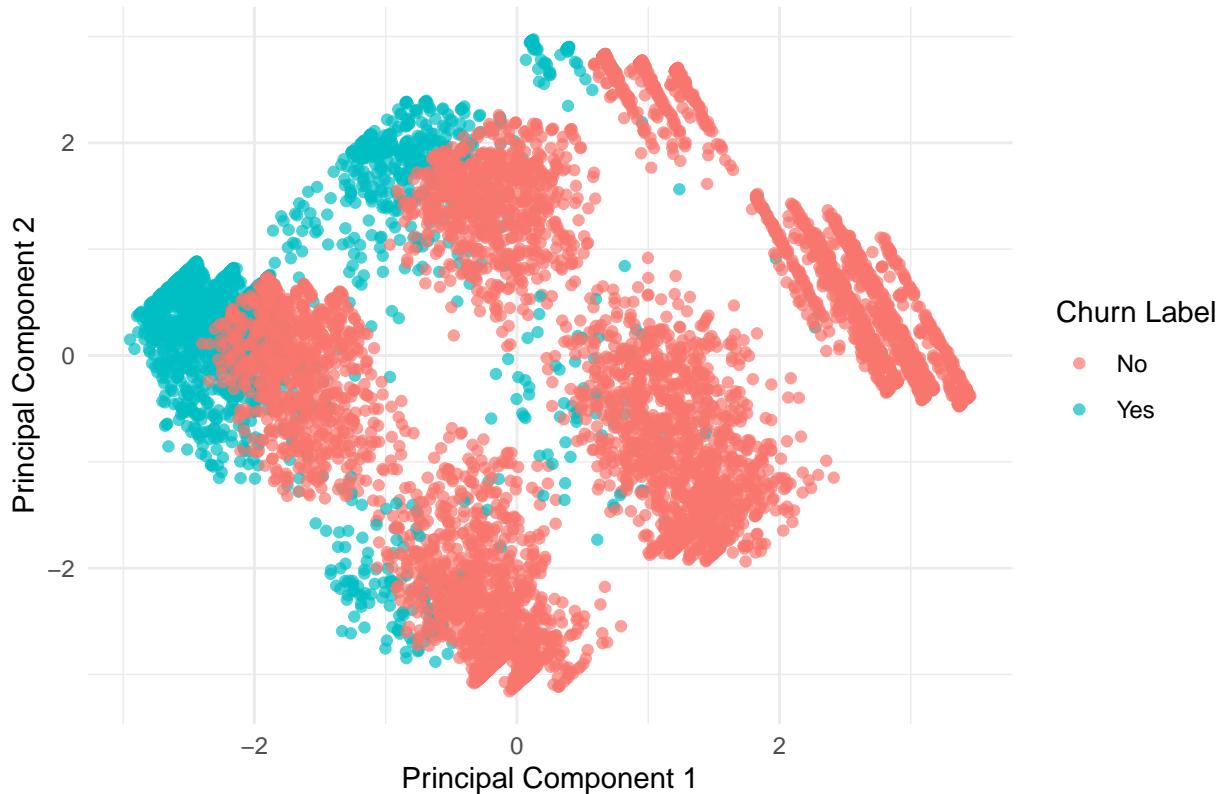
```
library(ggplot2)

# Create a dataframe with PC1 and PC2
pca_df <- as.data.frame(pca_result$x[, 1:2])
colnames(pca_df) <- c("PC1", "PC2")

# Add churn label for coloring
pca_df$ChurnLabel <- as.factor(telco_data$`Churn Label`)

# Plot
ggplot(pca_df, aes(x = PC1, y = PC2, color = ChurnLabel)) +
  geom_point(alpha = 0.7) +
  labs(title = "PCA 2D Visualization (PC1 vs PC2)",
       x = "Principal Component 1",
       y = "Principal Component 2",
       color = "Churn Label") +
  theme_minimal()
```

## PCA 2D Visualization (PC1 vs PC2)



```

library(plotly)
library(conflicted)

# Always prefer plotly's layout instead of graphics::layout
conflicts_prefer(plotly::layout)

## [conflicted] Will prefer plotly::layout over any other package.

# Extract first 3 principal components
pca_3d <- as.data.frame(pca_result$x[, 1:3])
colnames(pca_3d) <- c("PC1", "PC2", "PC3")
pca_3d$ChurnLabel <- as.factor(telco_data$`Churn Label`)

# Create 3D scatter plot
plot_ly(
  data = pca_3d,
  x = ~PC1, y = ~PC2, z = ~PC3,
  color = ~ChurnLabel,
  colors = c("steelblue", "tomato"),
  type = "scatter3d",
  mode = "markers",
  marker = list(size = 4, opacity = 0.7)
) %>%
  plotly::layout(
    title = "PCA 3D Visualization (PC1, PC2, PC3)",
    width = 800,
    height = 600
)
  
```

```

    scene = list(
      xaxis = list(title = "Principal Component 1"),
      yaxis = list(title = "Principal Component 2"),
      zaxis = list(title = "Principal Component 3")
    )
  )

library(umap)
library(ggplot2)

# Set UMAP configuration
umap_config <- umap.defaults
umap_config$n_neighbors <- 15
umap_config$min_dist <- 0.1
umap_config$random_state <- 42

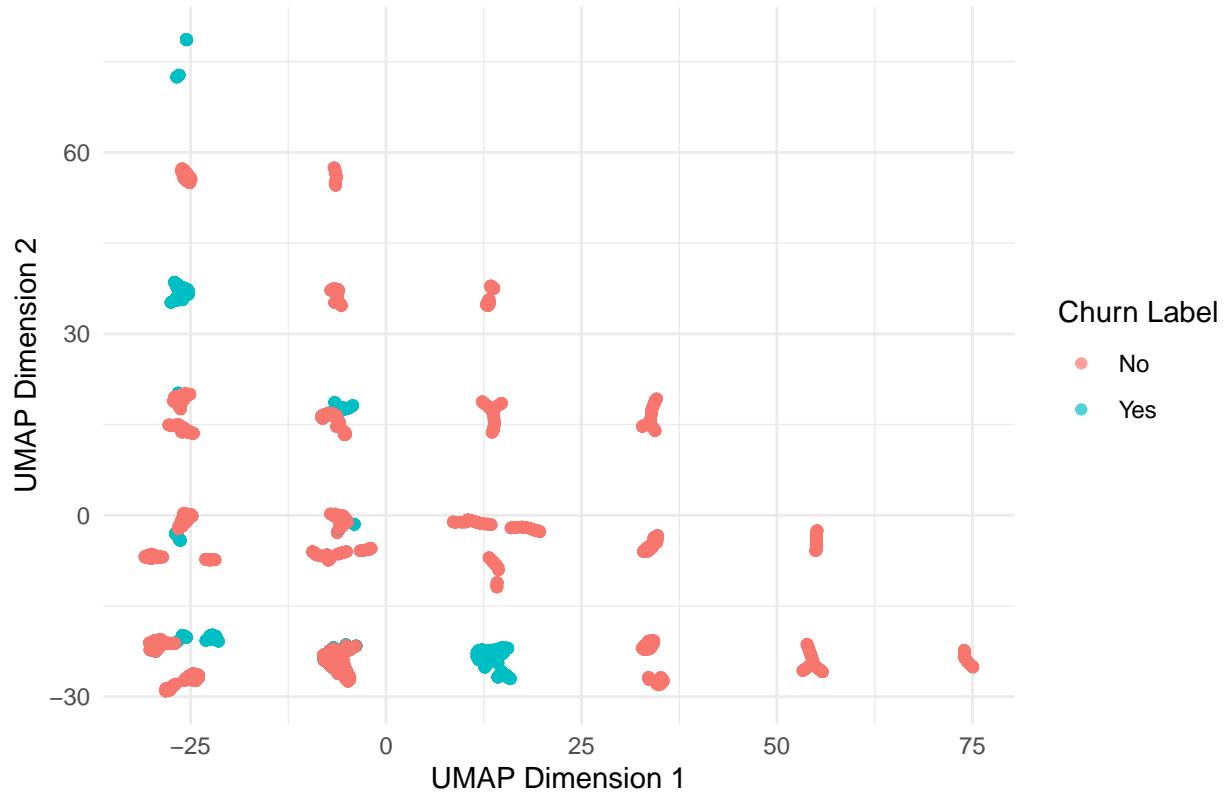
# Perform UMAP on the processed features
set.seed(42)
umap_result <- umap(as.matrix(processed_features), config = umap_config)

# Convert UMAP layout to dataframe
umap_df <- as.data.frame(umap_result$layout)
colnames(umap_df) <- c("UMAP1", "UMAP2")
umap_df$ChurnLabel <- as.factor(telco_data$`Churn Label`)

# Plot UMAP results
ggplot(umap_df, aes(x = UMAP1, y = UMAP2, color = ChurnLabel)) +
  geom_point(alpha = 0.7) +
  labs(title = "UMAP Visualization",
       x = "UMAP Dimension 1",
       y = "UMAP Dimension 2",
       color = "Churn Label") +
  theme_minimal()

```

## UMAP Visualization



```

library(Rtsne)
library(ggplot2)

# Remove duplicates from the matrix
processed_unique <- unique(as.matrix(processed_features))

duplicate_rows <- duplicated(as.data.frame(processed_features))
filtered_telco_data <- telco_data[!duplicate_rows, ]

# Run t-SNE
set.seed(42)
tsne_result <- Rtsne(
  processed_unique,
  dims = 2,
  perplexity = 30,
  max_iter = 500,
  verbose = TRUE
)

## Performing PCA
## Read the 6892 x 10 data matrix successfully!
## Using no_dims = 2, perplexity = 30.000000, and theta = 0.500000
## Computing input similarities...
## Building tree...
## Done in 0.20 seconds (sparsity = 0.015966)!
## Learning embedding...

```

```

## Iteration 50: error is 94.319461 (50 iterations in 0.51 seconds)
## Iteration 100: error is 70.580708 (50 iterations in 0.48 seconds)
## Iteration 150: error is 64.293336 (50 iterations in 0.45 seconds)
## Iteration 200: error is 61.406194 (50 iterations in 0.46 seconds)
## Iteration 250: error is 59.657152 (50 iterations in 0.45 seconds)
## Iteration 300: error is 1.844304 (50 iterations in 0.44 seconds)
## Iteration 350: error is 1.447246 (50 iterations in 0.44 seconds)
## Iteration 400: error is 1.181706 (50 iterations in 0.44 seconds)
## Iteration 450: error is 1.005213 (50 iterations in 0.45 seconds)
## Iteration 500: error is 0.884580 (50 iterations in 0.45 seconds)
## Fitting performed in 4.56 seconds.

```

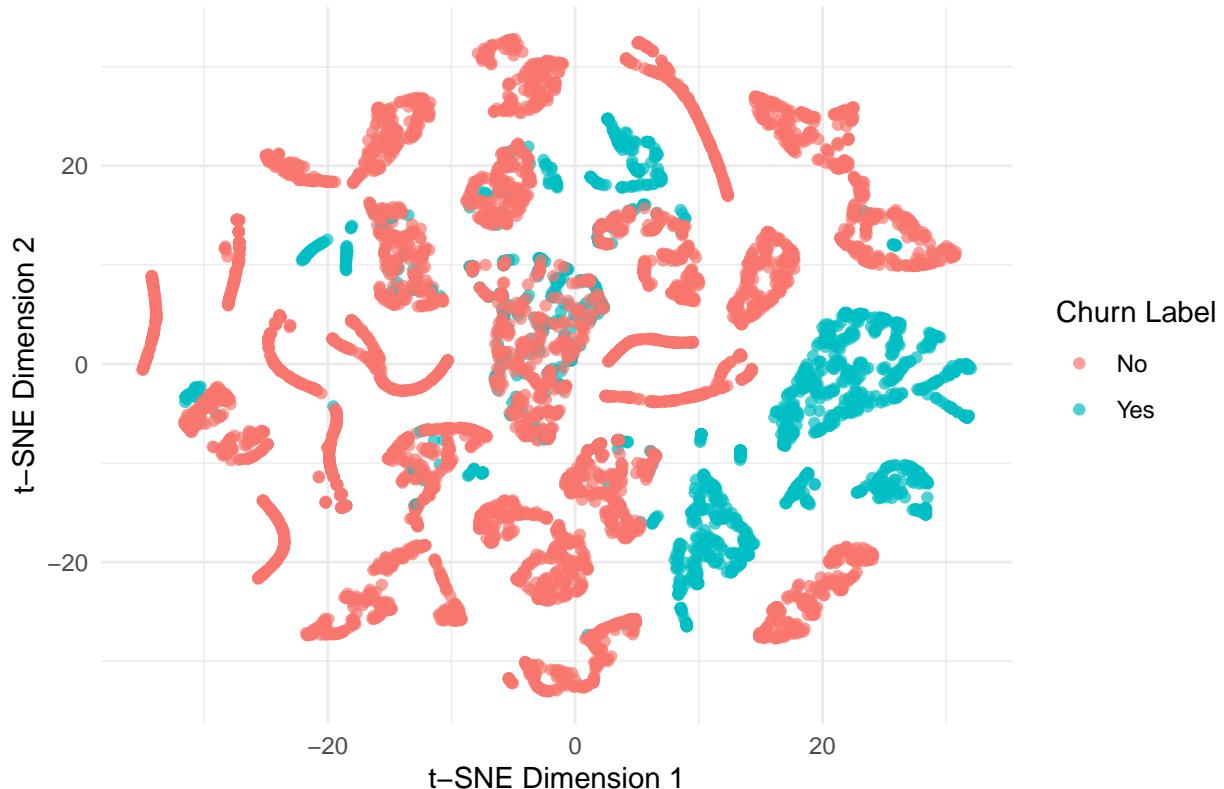
```

# Convert results to dataframe
tsne_df <- as.data.frame(tsne_result$Y)
colnames(tsne_df) <- c("TSNE1", "TSNE2")
tsne_df$ChurnLabel <- as.factor(filtered_telco_data$`Churn Label`)

# Plot t-SNE results
ggplot(tsne_df, aes(x = TSNE1, y = TSNE2, color = ChurnLabel)) +
  geom_point(alpha = 0.7) +
  labs(title = "t-SNE Visualization (Perplexity = 30)",
       x = "t-SNE Dimension 1",
       y = "t-SNE Dimension 2",
       color = "Churn Label") +
  theme_minimal()

```

t-SNE Visualization (Perplexity = 30)



```

library(Rtsne)
library(ggplot2)

# Remove duplicate rows from processed features
processed_unique <- unique(as.matrix(processed_features))

# Filter churn labels accordingly
duplicate_rows <- duplicated(as.data.frame(processed_features))
filtered_telco_data <- telco_data[!duplicate_rows, ]

# Define perplexity values
perplexities <- c(10, 30, 50)

# Loop through perplexity values and plot t-SNE
for (perp in perplexities) {
  set.seed(42)
  tsne_result <- Rtsne(processed_unique,
                        dims = 2,
                        perplexity = perp,
                        max_iter = 500,
                        verbose = FALSE)

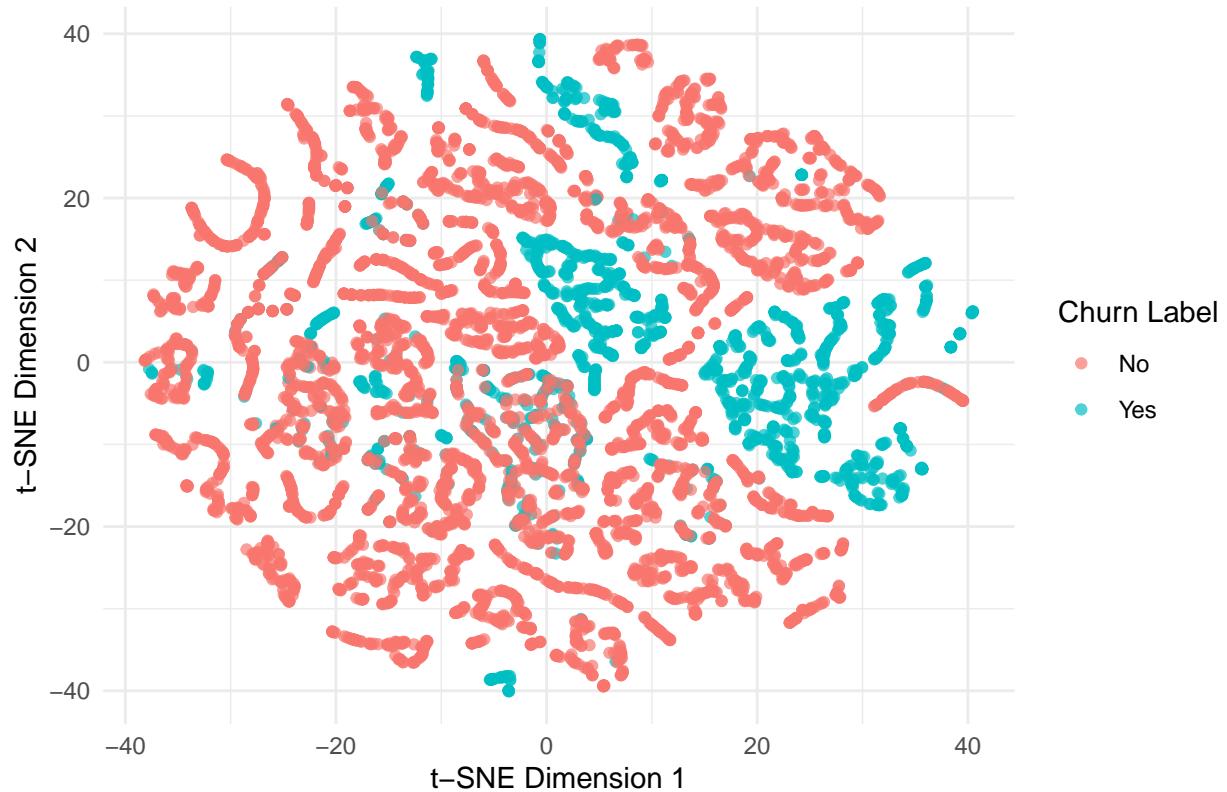
  tsne_df <- as.data.frame(tsne_result$Y)
  colnames(tsne_df) <- c("TSNE1", "TSNE2")
  tsne_df$ChurnLabel <- as.factor(filtered_telco_data$`Churn Label`)

  p <- ggplot(tsne_df, aes(x = TSNE1, y = TSNE2, color = ChurnLabel)) +
    geom_point(alpha = 0.7) +
    labs(
      title = paste("t-SNE Visualization (Perplexity =", perp, ")"),
      x = "t-SNE Dimension 1",
      y = "t-SNE Dimension 2",
      color = "Churn Label"
    ) +
    theme_minimal()

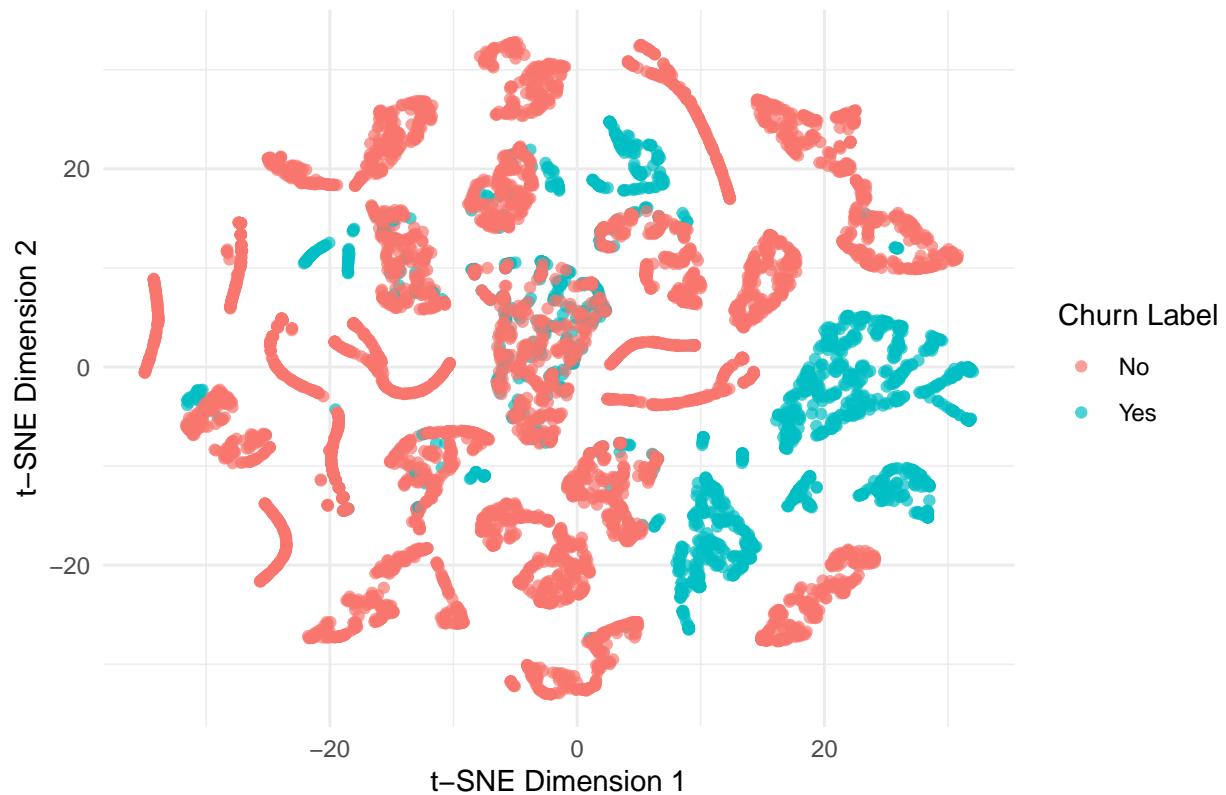
  print(p)
}

```

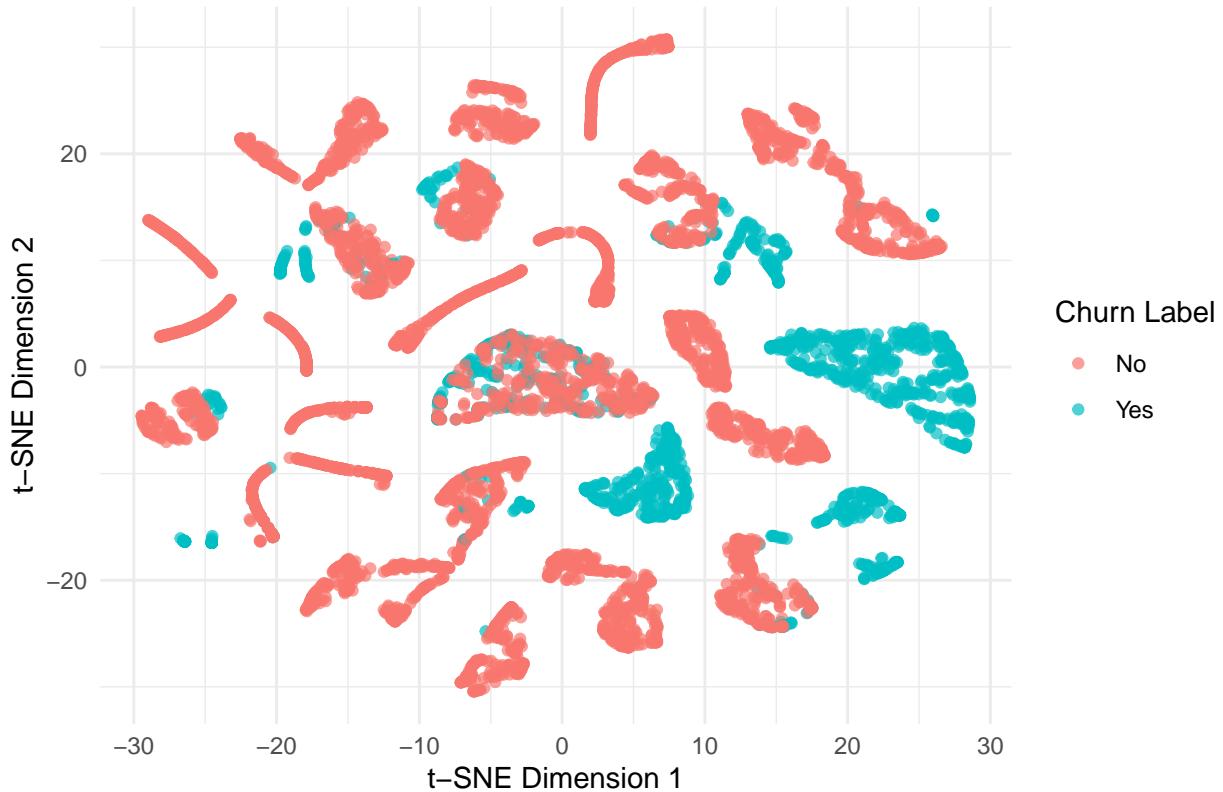
t-SNE Visualization (Perplexity = 10 )



t-SNE Visualization (Perplexity = 30 )



## t-SNE Visualization (Perplexity = 50 )



```

library(umap)
library(ggplot2)

# Define combinations of parameters
n_neighbors_list <- c(10, 30, 50)
min_dist_list <- c(0.1, 0.5)

# Loop through each combination of n_neighbors and min_dist
for (n_neighbors in n_neighbors_list) {
  for (min_dist in min_dist_list) {

    # Configure UMAP parameters
    umap_config <- umap.defaults
    umap_config$n_neighbors <- n_neighbors
    umap_config$min_dist <- min_dist
    umap_config$random_state <- 42

    # Run UMAP
    set.seed(42)
    umap_result <- umap(as.matrix(processed_features), config = umap_config)

    # Create data frame for plotting
    umap_df <- as.data.frame(umap_result$layout)
    colnames(umap_df) <- c("UMAP1", "UMAP2")
    umap_df$ChurnLabel <- as.factor(telco_data$`Churn Label`)
  }
}

```

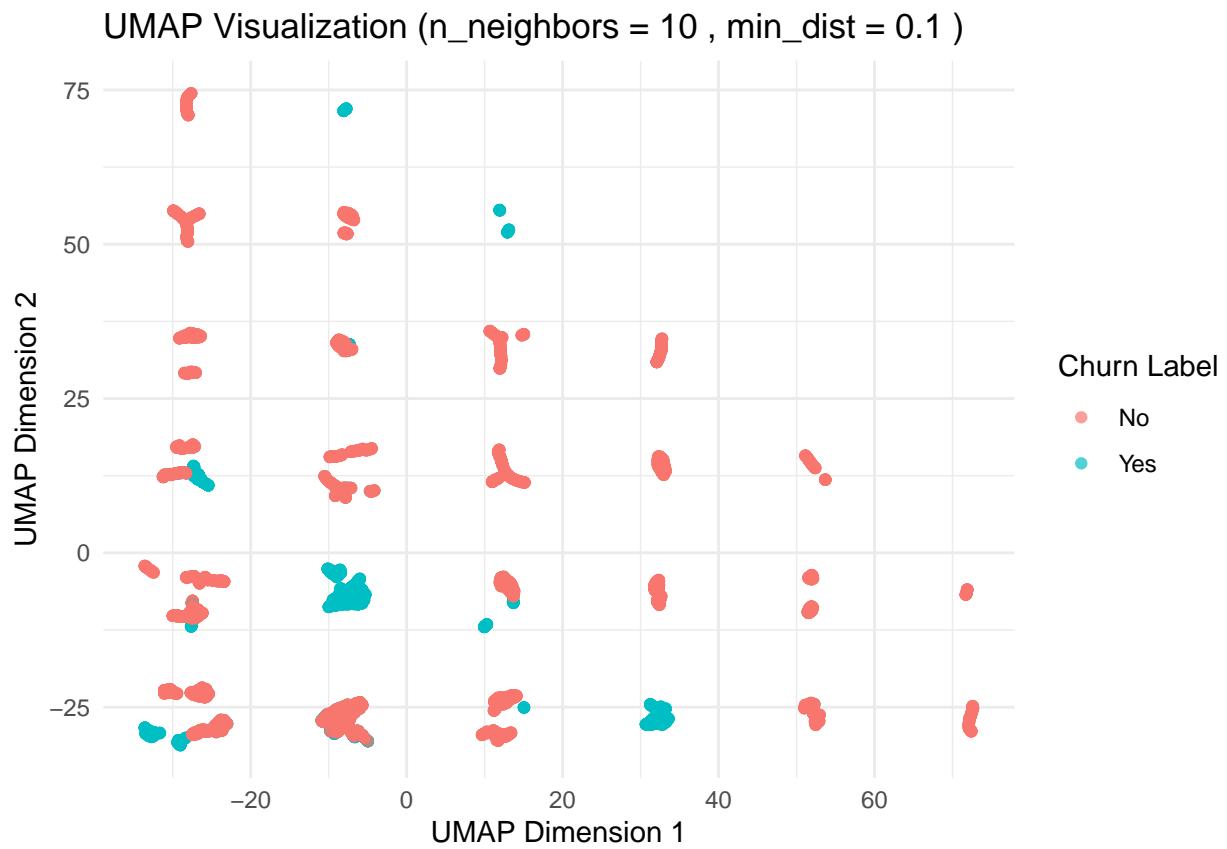
```

# Create and show the plot
plot_title <- paste("UMAP Visualization (n_neighbors =", n_neighbors, ", min_dist =", min_dist, ")")

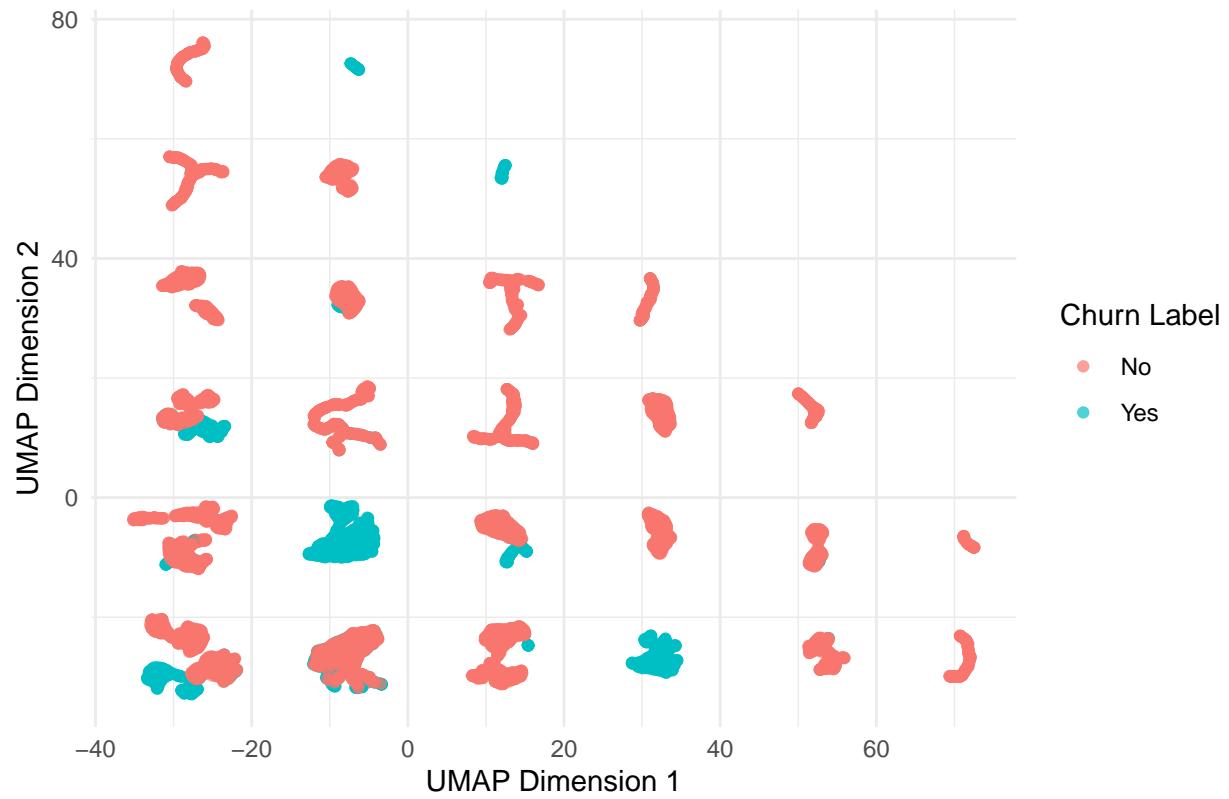
p <- ggplot(umap_df, aes(x = UMAP1, y = UMAP2, color = ChurnLabel)) +
  geom_point(alpha = 0.7) +
  labs(title = plot_title, x = "UMAP Dimension 1", y = "UMAP Dimension 2", color = "Churn Label") +
  theme_minimal()

print(p)
}
}

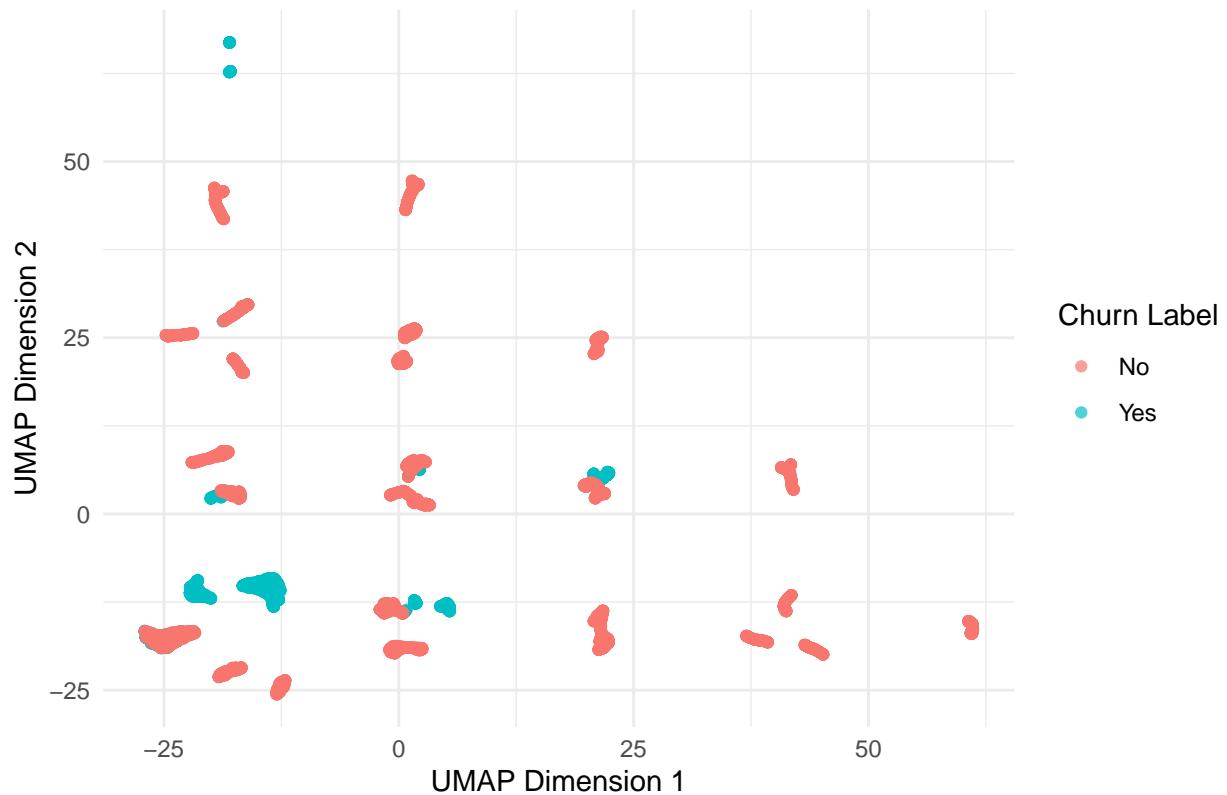
```



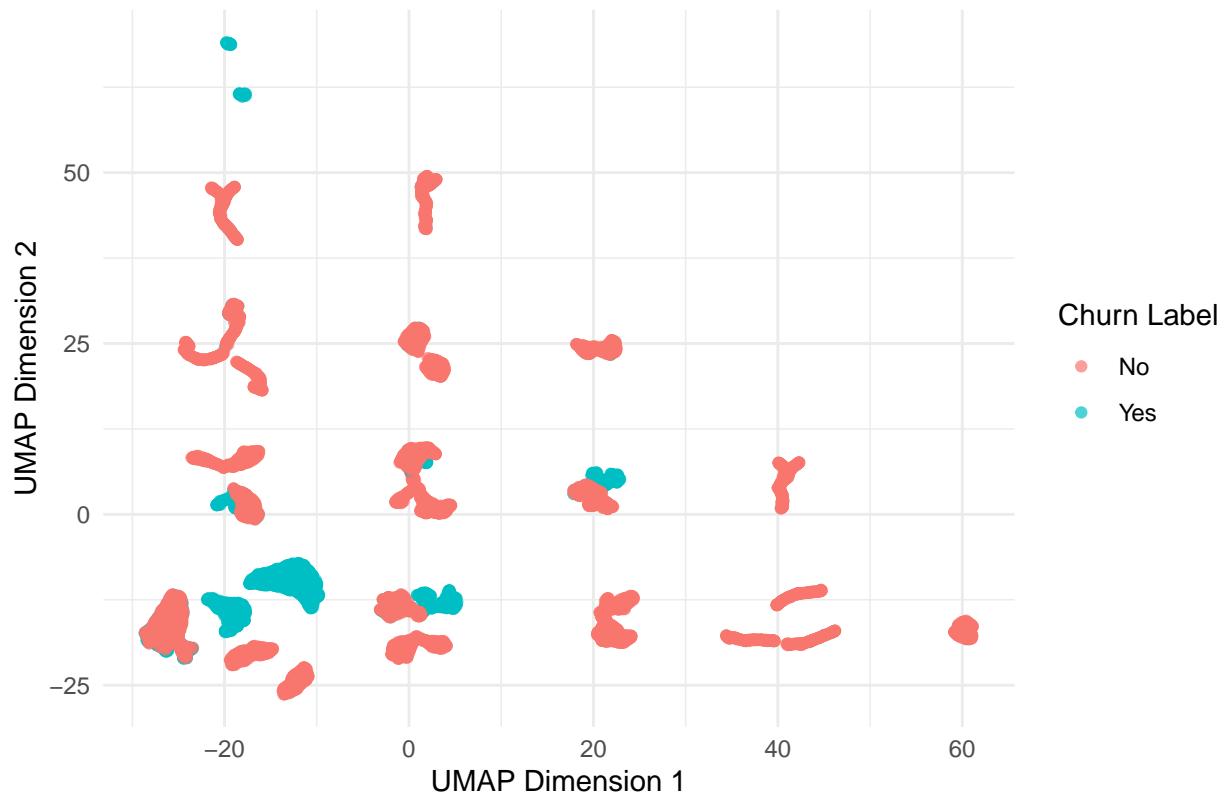
UMAP Visualization (n\_neighbors = 10 , min\_dist = 0.5 )



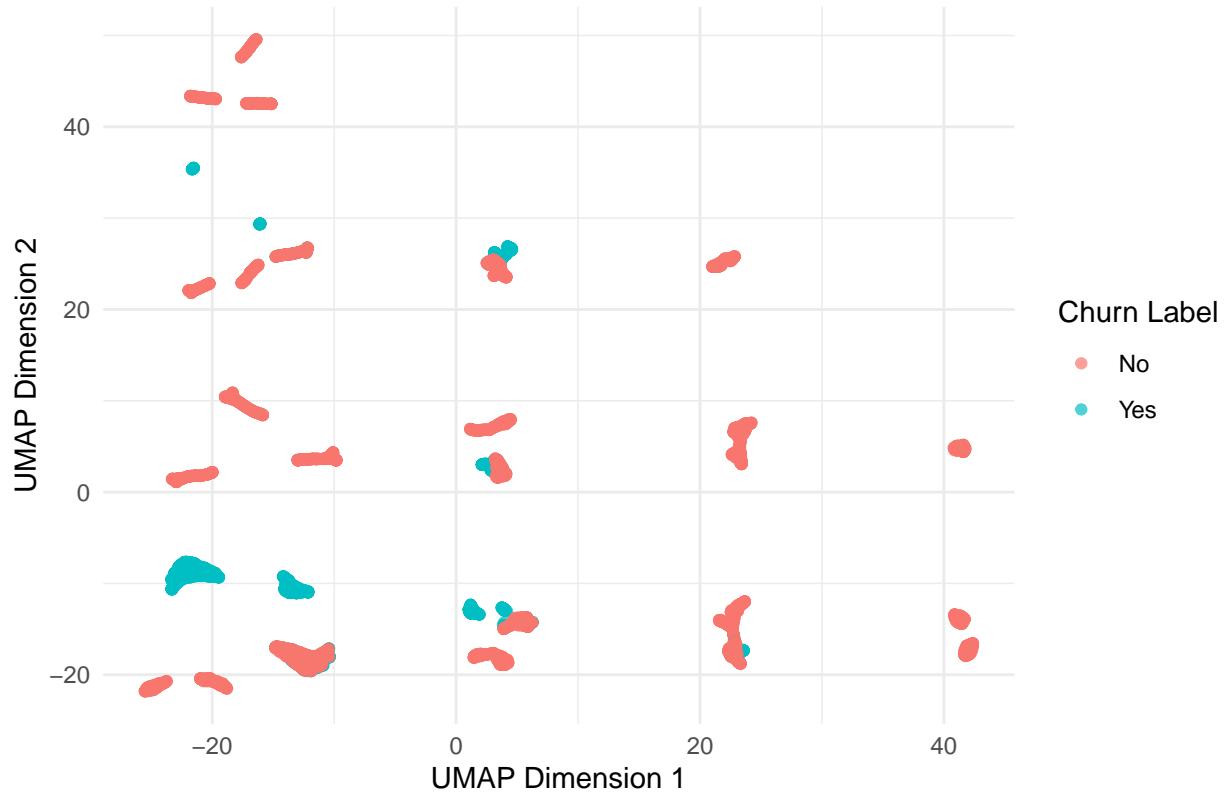
UMAP Visualization (n\_neighbors = 30 , min\_dist = 0.1 )



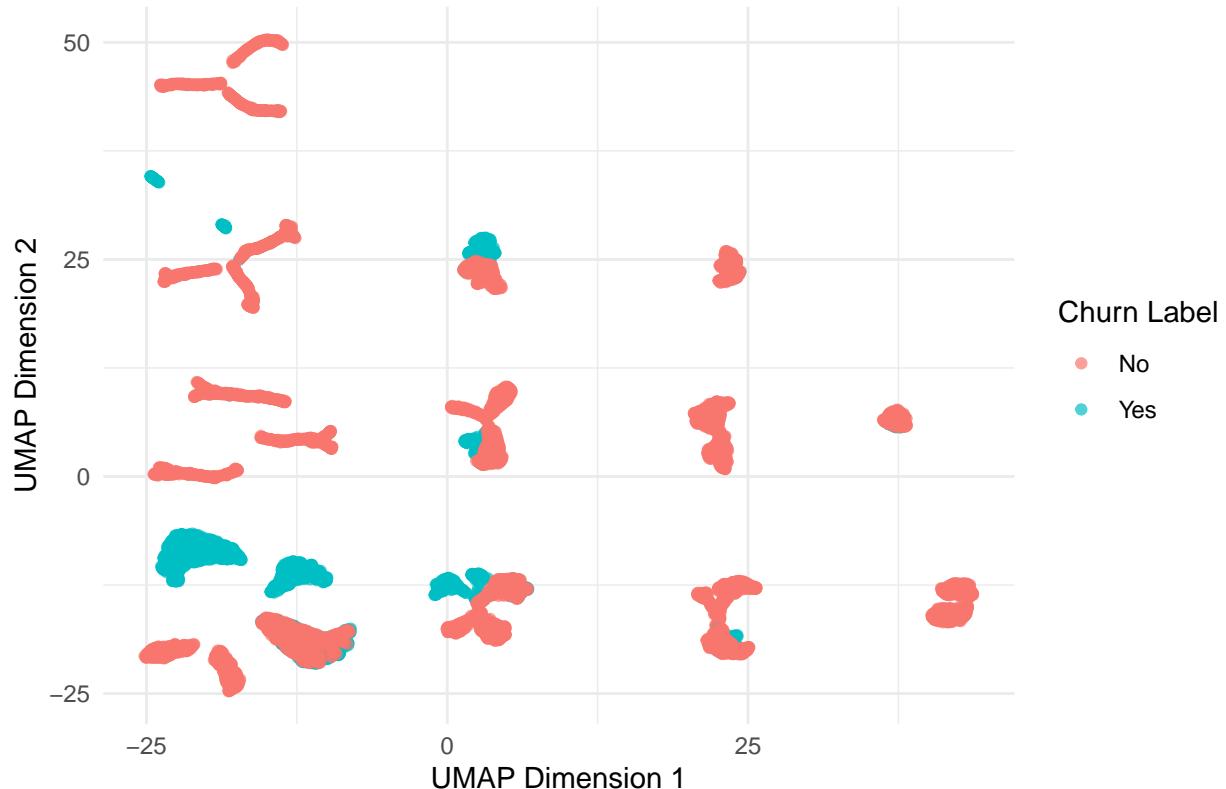
UMAP Visualization (n\_neighbors = 30 , min\_dist = 0.5 )



UMAP Visualization (n\_neighbors = 50 , min\_dist = 0.1 )



## UMAP Visualization (n\_neighbors = 50 , min\_dist = 0.5 )



```

library(dplyr)

# Create interaction terms
telco_data <- telco_data %>%
  mutate(
    Tenure_Charge = `Tenure in Months` * `Monthly Charge`,
    Satisfaction_Contract = `Satisfaction Score` * as.numeric(as.factor(Contract)),

    # Transform numerical features
    Log_Monthly_Charge = log1p(`Monthly Charge`),
    Square_Satisfaction = `Satisfaction Score`^2,

    # Create new features
    Avg_Charge_Per_Month = `Total Charges` / ifelse(`Tenure in Months` == 0, 1, `Tenure in Months`),
    High_Spender = as.integer(`Monthly Charge` > median(`Monthly Charge`, na.rm = TRUE))
  )

categorical_features <- c("Contract", "Internet Type")

numerical_features <- c(
  "Tenure in Months", "Monthly Charge", "Satisfaction Score",
  "Tenure_Charge", "Satisfaction_Contract", "Log_Monthly_Charge",
  "Square_Satisfaction", "Avg_Charge_Per_Month"
)

```

```

library(recipes)
library(conflicted)
conflicts_prefer(dplyr::select)

## [conflicted] Will prefer dplyr::select over any other package.

# Subset relevant columns
telco_subset <- telco_data %>%
  select(all_of(c(numerical_features, categorical_features)))

# Define preprocessing pipeline
rec <- recipe(~ ., data = telco_subset) %>%
  step_dummy(all_nominal_predictors(), one_hot = TRUE) %>%
  step_center(all_numeric_predictors()) %>%
  step_scale(all_numeric_predictors())

# Prepare and apply transformations
prepped_data <- prep(rec, training = telco_subset)
processed_features <- bake(prepped_data, new_data = telco_subset)

library(Rtsne)
library(ggplot2)

# Step 1: Convert to data frame for de-duplication
processed_df <- as.data.frame(processed_features)

# Step 2: Identify and remove duplicate rows
non_duplicate_mask <- !duplicated(processed_df)
processed_unique <- processed_df[non_duplicate_mask, ]

# Step 3: Filter labels to match non-duplicated rows
filtered_labels <- telco_data$`Churn Label`[non_duplicate_mask]

# Step 4: Convert to matrix for Rtsne
processed_matrix <- as.matrix(processed_unique)

# Step 5: Run t-SNE
set.seed(42)
tsne_result <- Rtsne(processed_matrix,
                      dims = 2,
                      perplexity = 30,
                      max_iter = 500,
                      verbose = FALSE)

# Step 6: Prepare DataFrame for plotting
tsne_df <- as.data.frame(tsne_result$Y)
colnames(tsne_df) <- c("TSNE1", "TSNE2")
tsne_df$ChurnLabel <- as.factor(filtered_labels)

# Step 7: Plot
ggplot(tsne_df, aes(x = TSNE1, y = TSNE2, color = ChurnLabel)) +
  geom_point(alpha = 0.7) +
  labs(title = "t-SNE with Engineered Features (No Duplicates)",

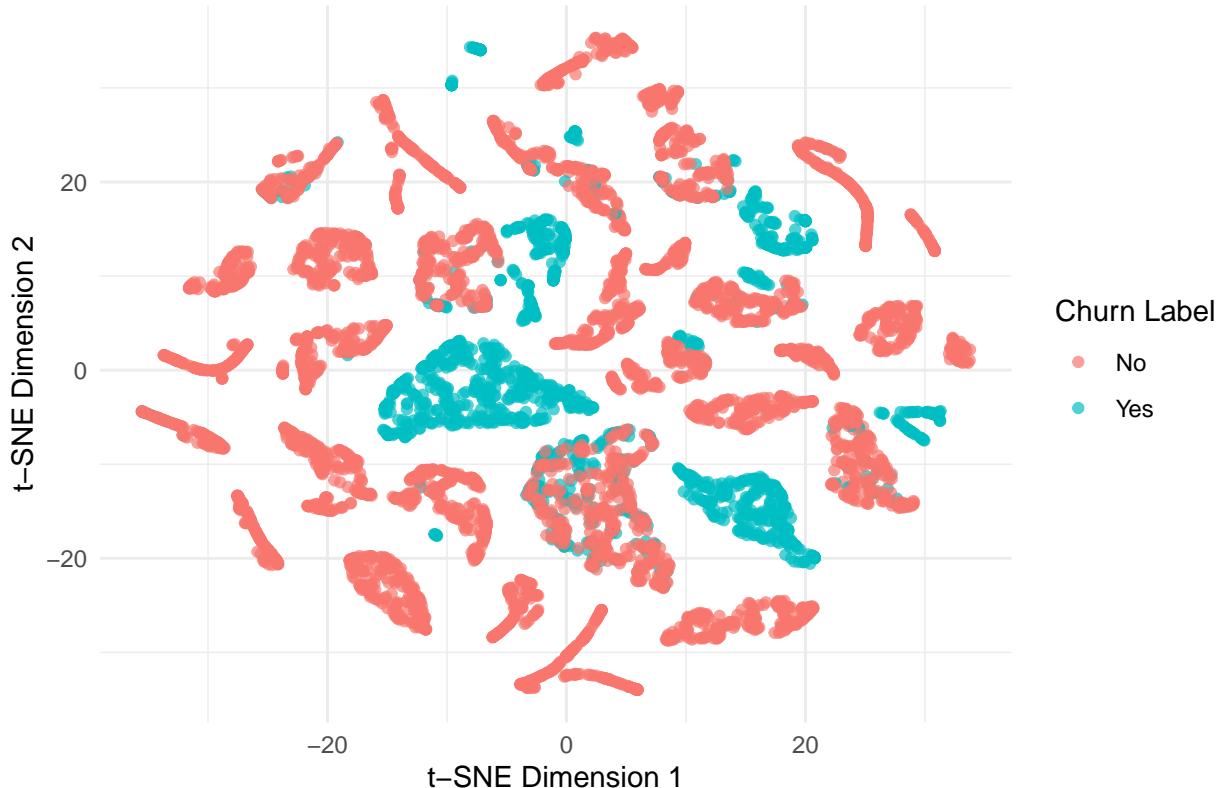
```

```

x = "t-SNE Dimension 1",
y = "t-SNE Dimension 2",
color = "Churn Label") +
theme_minimal()

```

t-SNE with Engineered Features (No Duplicates)



```

library(umap)
library(ggplot2)

# Configure UMAP
umap_config <- umap.defaults
umap_config$n_neighbors <- 30
umap_config$min_dist <- 0.1
umap_config$random_state <- 42

# Run UMAP
set.seed(42)
umap_result <- umap(as.matrix(processed_features), config = umap_config)

# Create a dataframe for plotting
umap_df <- as.data.frame(umap_result$layout)
colnames(umap_df) <- c("UMAP1", "UMAP2")
umap_df$ChurnLabel <- as.factor(telco_data$`Churn Label`)

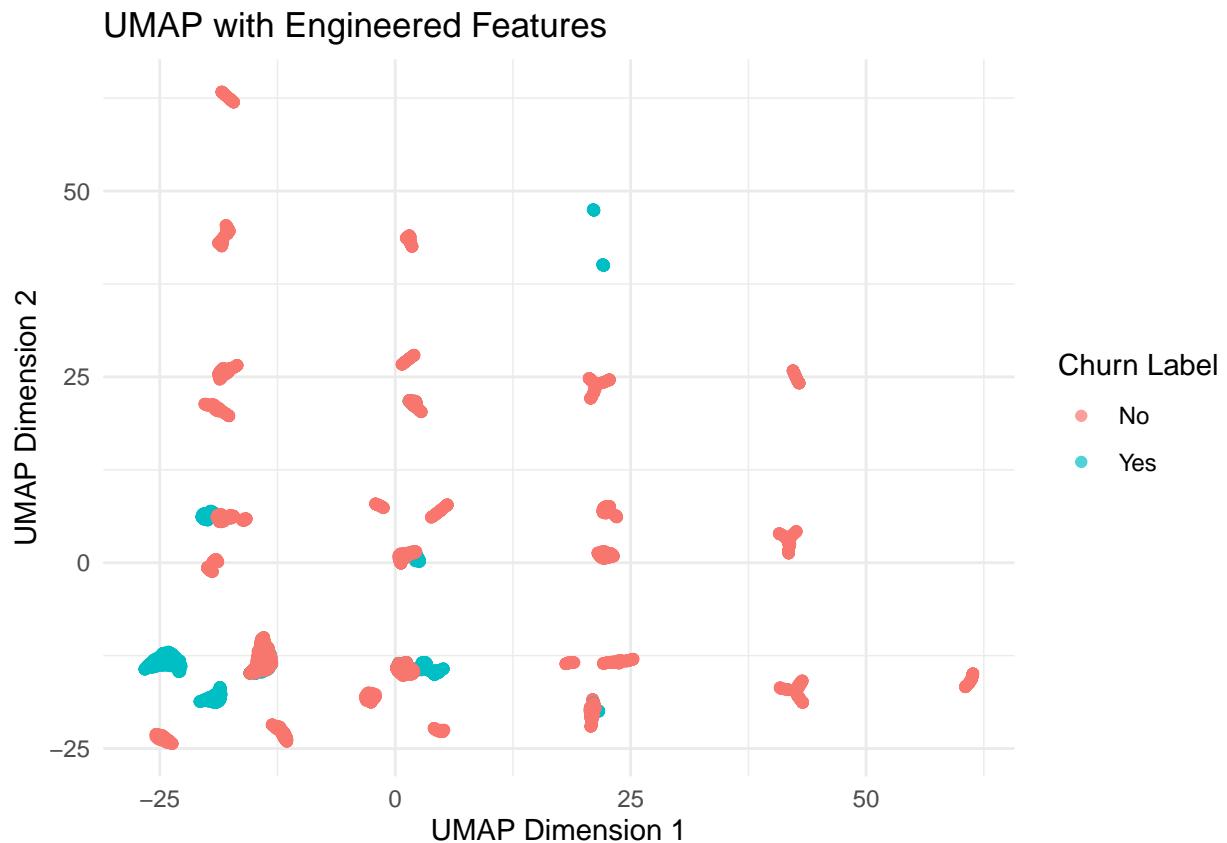
# Plot the UMAP result
ggplot(umap_df, aes(x = UMAP1, y = UMAP2, color = ChurnLabel)) +

```

```

geom_point(alpha = 0.7) +
labs(title = "UMAP with Engineered Features",
x = "UMAP Dimension 1",
y = "UMAP Dimension 2",
color = "Churn Label") +
theme_minimal()

```



```

library(caret)
library(randomForest)
library(dplyr)

# Encode target variable ('Yes' -> 1, 'No' -> 0)
y <- as.factor(ifelse(telco_data$`Churn Label` == "Yes", 1, 0))

colnames(processed_features) <- make.names(paste0("Feature_", seq_len(ncol(processed_features))))

# Combine features and target
data_model <- as.data.frame(processed_features)
data_model$Churn <- y

# Set up stratified 5-fold cross-validation
set.seed(42)
cv_folds <- createFolds(data_model$Churn, k = 5, list = TRUE, returnTrain = TRUE)

# Initialize metric lists

```

```

accuracy <- c()
precision <- c()
recall <- c()
f1 <- c()

# Cross-validation loop
for (i in 1:5) {
  train_indices <- cv_folds[[i]]
  train_data <- data_model[train_indices, ]
  test_data <- data_model[-train_indices, ]

  model <- randomForest(Churn ~ ., data = train_data, ntree = 100, importance = TRUE)
  predictions <- predict(model, test_data)
  truth <- test_data$Churn

  # Confusion matrix
  cm <- confusionMatrix(predictions, truth, positive = "1")

  # Store metrics
  accuracy <- c(accuracy, cm$overall["Accuracy"])
  precision <- c(precision, cm$byClass["Precision"])
  recall <- c(recall, cm$byClass["Recall"])
  f1 <- c(f1, cm$byClass["F1"])
}

```

```

# Print aggregated results
cat("Random Forest Classifier Results:\n")

```

```

## Random Forest Classifier Results:

cat("K-Fold Cross-Validation Results:\n")

```

```

## K-Fold Cross-Validation Results:

cat(sprintf("Accuracy: %.4f\n", mean(accuracy)))

```

```

## Accuracy: 0.9487

```

```

cat(sprintf("Precision: %.4f\n", mean(precision)))

```

```

## Precision: 0.9628

```

```

cat(sprintf("Recall: %.4f\n", mean(recall)))

```

```

## Recall: 0.8395

```

```

cat(sprintf("F1-Score: %.4f\n", mean(f1)))

```

```

## F1-Score: 0.8968

```

```

library(caret)
library(ranger)
library(dplyr)
library(conflicted)
conflicts_prefer(dplyr::filter)

## [conflicted] Will prefer dplyr::filter over any other package.

conflicts_prefer(dplyr::select) # also for select if needed

## [conflicted] Removing existing preference.
## [conflicted] Will prefer dplyr::select over any other package.

# Encode the target variable
y <- as.factor(ifelse(telco_data$`Churn Label` == "Yes", 1, 0))

# Rename features to be safe for modeling
colnames(processed_features) <- make.names(paste0("Feature_", seq_len(ncol(processed_features)))))

# Combine features and target
data_model <- as.data.frame(processed_features)
data_model$Churn <- factor(y, labels = c("No", "Yes"))

# Set up stratified 5-fold cross-validation with ROC
train_control <- trainControl(
  method = "cv",
  number = 5,
  classProbs = TRUE,
  summaryFunction = twoClassSummary,
  savePredictions = "final"
)

# Define parameter grid
rf_grid <- expand.grid(
  .mtry = c(2, 3, 4),
  .splitrule = "gini",
  .min.node.size = c(1, 2, 4)
)

# Train the model using caret + ranger
set.seed(42)
rf_model <- train(
  Churn ~.,
  data = data_model,
  method = "ranger",
  trControl = train_control,
  metric = "ROC",
  tuneGrid = rf_grid,
  importance = 'impurity'
)

# Print best parameters
cat("Best Parameters:\n")

```

```

## Best Parameters:

print(rf_model$bestTune)

##   mtry splitrule min.node.size
## 5     3       gini           2

# Get final predictions and compute evaluation metrics manually
preds <- rf_model$pred
best_params <- rf_model$bestTune

# Filter predictions to the best parameter combination
preds_best <- preds %>%
  filter(
    mtry == best_params$mtry,
    splitrule == best_params$splitrule,
    min.node.size == best_params$min.node.size
  )

# Compute confusion matrix
cm <- confusionMatrix(preds_best$pred, preds_best$obs, positive = "Yes")

# Output metrics
cat("\nTuned Random Forest Classifier Results (Cross-Validation):\n")

##  

## Tuned Random Forest Classifier Results (Cross-Validation):  
  

cat(sprintf("Accuracy: %.4f\n", cm$overall["Accuracy"]))

## Accuracy: 0.9497  
  

cat(sprintf("Precision: %.4f\n", cm$byClass["Precision"]))

## Precision: 0.9610  
  

cat(sprintf("Recall: %.4f\n", cm$byClass["Recall"]))

## Recall: 0.8448  
  

cat(sprintf("F1-Score: %.4f\n", cm$byClass["F1"]))

## F1-Score: 0.8992  
  

library(caret)
library(dplyr)

# Encode target as factor: "Yes" = 1, "No" = 0
y <- as.factor(ifelse(telco_data$`Churn Label` == "Yes", 1, 0))

```

```

# Combine features and label
data_model <- processed_features
data_model$Churn <- y

data_model$Churn <- factor(data_model$Churn, labels = c("No", "Yes"))

# Set up stratified 5-fold cross-validation
set.seed(42)
cv_folds <- createFolds(data_model$Churn, k = 5, list = TRUE, returnTrain = TRUE)

# Initialize metric lists
accuracy_lr <- c()
precision_lr <- c()
recall_lr <- c()
f1_lr <- c()

# Run manual cross-validation loop
for (i in 1:5) {
  train_idx <- cv_folds[[i]]
  train_data <- data_model[train_idx, ]
  test_data <- data_model[-train_idx, ]

  # Fit logistic regression
  log_model <- glm(Churn ~ ., data = train_data, family = binomial)

  # Predict probabilities and class labels
  probs <- predict(log_model, test_data, type = "response")
  preds <- ifelse(probs > 0.5, "Yes", "No") %>% as.factor()

  # Evaluation
  cm <- confusionMatrix(preds, test_data$Churn, positive = "Yes")
  accuracy_lr <- c(accuracy_lr, cm$overall["Accuracy"])
  precision_lr <- c(precision_lr, cm$byClass["Precision"])
  recall_lr <- c(recall_lr, cm$byClass["Recall"])
  f1_lr <- c(f1_lr, cm$byClass["F1"])
}

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

```

```

# Print averaged results
cat("Logistic Regression Results:\n")

```

```

## Logistic Regression Results:

```

```

cat(sprintf("Accuracy: %.4f\n", mean(accuracy_lr)))

```

```

## Accuracy: 0.9473

```

```

cat(sprintf("Precision: %.4f\n", mean(precision_lr)))

## Precision: 0.9513

cat(sprintf("Recall: %.4f\n", mean(recall_lr)))

## Recall: 0.8448

cat(sprintf("F1-Score: %.4f\n", mean(f1_lr)))

## F1-Score: 0.8948

library(caret)
library(dplyr)

# Encode target
y <- as.factor(ifelse(telco_data$`Churn Label` == "Yes", 1, 0))

# Rename features to valid names
colnames(processed_features) <- make.names(paste0("Feature_", seq_len(ncol(processed_features))))

# Combine features and target
data_model <- as.data.frame(processed_features)
data_model$Churn <- factor(y, labels = c("No", "Yes"))

# Define 5-fold stratified cross-validation
set.seed(42)
train_control <- trainControl(
  method = "cv",
  number = 5,
  classProbs = TRUE,
  summaryFunction = twoClassSummary,
  savePredictions = "final"
)

# Define parameter grid for glmnet
lambda_values <- 1 / c(0.01, 0.1, 1, 10, 100)

# Train logistic regression with regularization
set.seed(42)
log_model <- train(
  Churn ~.,
  data = data_model,
  method = "glmnet",
  trControl = train_control,
  metric = "ROC",
  tuneGrid = expand.grid(
    alpha = 0,
    lambda = lambda_values
  ),
  family = "binomial"
)

```

```

)

# Show best parameters
cat("Best Parameters:\n")

## Best Parameters:

print(log_model$bestTune)

##   alpha lambda
## 1     0    0.01

# Extract predictions for best model
best_params <- log_model$bestTune
preds_best <- log_model$pred %>%
  filter(
    alpha == best_params$alpha,
    lambda == best_params$lambda
  )

# Compute confusion matrix
cm <- confusionMatrix(preds_best$pred, preds_best$obs, positive = "Yes")

# Output final metrics
cat("\nTuned Logistic Regression Results (Cross-Validation):\n")

## Tuned Logistic Regression Results (Cross-Validation):

cat(sprintf("Accuracy: %.4f\n", cm$overall["Accuracy"]))

## Accuracy: 0.9429

cat(sprintf("Precision: %.4f\n", cm$byClass["Precision"]))

## Precision: 0.9437

cat(sprintf("Recall: %.4f\n", cm$byClass["Recall"]))

## Recall: 0.8347

cat(sprintf("F1-Score: %.4f\n", cm$byClass["F1"]))

## F1-Score: 0.8859

```

```

library(xgboost)
library(caret)
library(dplyr)

# Encode target variable: "Yes" = 1, "No" = 0
y <- ifelse(telco_data$`Churn Label` == "Yes", 1, 0)

# Convert processed_features and y to matrices (xgboost requires matrices)
X <- as.matrix(processed_features)
y <- as.numeric(y)

# Stratified 5-fold split
set.seed(42)
folds <- createFolds(y, k = 5, list = TRUE, returnTrain = TRUE)

# Initialize metric holders
accuracy_xgb <- c()
precision_xgb <- c()
recall_xgb <- c()
f1_xgb <- c()

# Cross-validation with early stopping
for (i in 1:5) {
  train_idx <- folds[[i]]
  test_idx <- setdiff(1:nrow(X), train_idx)

  # Training/validation split for early stopping
  train_X_full <- X[train_idx, ]
  train_y_full <- y[train_idx]

  n <- floor(0.8 * length(train_y_full))
  train_X <- train_X_full[1:n, ]
  train_y <- train_y_full[1:n]
  val_X <- train_X_full[(n + 1):length(train_y_full), ]
  val_y <- train_y_full[(n + 1):length(train_y_full)]

  # DMatrix objects for XGBoost
  dtrain <- xgb.DMatrix(data = train_X, label = train_y)
  dval <- xgb.DMatrix(data = val_X, label = val_y)
  dtest <- xgb.DMatrix(data = X[test_idx, ])

  # Train XGBoost with early stopping
  xgb_model <- xgb.train(
    params = list(
      objective = "binary:logistic",
      eval_metric = "logloss",
      max_depth = 6,
      eta = 0.3
    ),
    data = dtrain,
    nrounds = 100,
    watchlist = list(val = dval),
    early_stopping_rounds = 10,

```

```

    verbose = 0
)

# Predict and evaluate
preds <- predict(xgb_model, dtest)
preds_class <- ifelse(preds > 0.5, 1, 0)
true_labels <- y[test_idx]

cm <- confusionMatrix(
  factor(preds_class, levels = c(0, 1)),
  factor(true_labels, levels = c(0, 1)),
  positive = "1"
)

accuracy_xgb <- c(accuracy_xgb, cm$overall["Accuracy"])
precision_xgb <- c(precision_xgb, cm$byClass["Precision"])
recall_xgb <- c(recall_xgb, cm$byClass["Recall"])
f1_xgb <- c(f1_xgb, cm$byClass["F1"])
}

# Print averaged results
cat("XGBoost Results with Early Stopping:\n")

## XGBoost Results with Early Stopping:

cat(sprintf("Accuracy: %.4f\n", mean(accuracy_xgb)))

## Accuracy: 0.9465

cat(sprintf("Precision: %.4f\n", mean(precision_xgb)))

## Precision: 0.9216

cat(sprintf("Recall: %.4f\n", mean(recall_xgb)))

## Recall: 0.8727

cat(sprintf("F1-Score: %.4f\n", mean(f1_xgb)))

## F1-Score: 0.8964

library(caret)
library(xgboost)
library(dplyr)

# Encode target variable: "Yes" = 1, "No" = 0
y <- ifelse(telco_data$`Churn Label` == "Yes", 1, 0)
y <- factor(y, levels = c(0, 1), labels = c("No", "Yes"))

```

```

# Combine features and label
data_model <- processed_features
data_model$Churn <- y

set.seed(42)
train_control <- trainControl(
  method = "cv",
  number = 5,
  summaryFunction = defaultSummary,
  savePredictions = TRUE,
  verboseIter = TRUE
)

# Keep original grid
xgb_grid <- expand.grid(
  nrounds = c(50, 100, 200),
  max_depth = c(3, 5, 7),
  eta = c(0.01, 0.1, 0.2),
  gamma = c(0, 1, 5),
  colsample_bytree = c(0.8, 1.0),
  min_child_weight = 1,
  subsample = c(0.8, 1.0)
)

set.seed(42)
xgb_model <- train(
  Churn ~.,
  data = data_model,
  method = "xgbTree",
  trControl = train_control,
  tuneGrid = xgb_grid,
  metric = "Accuracy"
)

## + Fold1: eta=0.01, max_depth=3, gamma=0, colsample_bytree=0.8, min_child_weight=1, subsample=0.8, nro
## [11:58:12] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [11:58:12] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## - Fold1: eta=0.01, max_depth=3, gamma=0, colsample_bytree=0.8, min_child_weight=1, subsample=0.8, nro
## + Fold1: eta=0.01, max_depth=3, gamma=0, colsample_bytree=0.8, min_child_weight=1, subsample=1.0, nro
## [11:58:12] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [11:58:12] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## - Fold1: eta=0.01, max_depth=3, gamma=0, colsample_bytree=0.8, min_child_weight=1, subsample=1.0, nro
## + Fold1: eta=0.01, max_depth=3, gamma=0, colsample_bytree=1.0, min_child_weight=1, subsample=0.8, nro
## [11:58:13] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [11:58:13] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## - Fold1: eta=0.01, max_depth=3, gamma=0, colsample_bytree=1.0, min_child_weight=1, subsample=0.8, nro
## + Fold1: eta=0.01, max_depth=3, gamma=0, colsample_bytree=1.0, min_child_weight=1, subsample=1.0, nro
## [11:58:13] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [11:58:13] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## - Fold1: eta=0.01, max_depth=3, gamma=0, colsample_bytree=1.0, min_child_weight=1, subsample=1.0, nro
## + Fold1: eta=0.01, max_depth=3, gamma=1, colsample_bytree=0.8, min_child_weight=1, subsample=0.8, nro
## [11:58:13] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [11:58:13] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead

```















































































```

## [12:02:24] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [12:02:24] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## - Fold5: eta=0.20, max_depth=7, gamma=0, colsample_bytree=1.0, min_child_weight=1, subsample=1.0, nrof
## + Fold5: eta=0.20, max_depth=7, gamma=1, colsample_bytree=0.8, min_child_weight=1, subsample=0.8, nrof
## [12:02:25] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [12:02:25] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## - Fold5: eta=0.20, max_depth=7, gamma=1, colsample_bytree=0.8, min_child_weight=1, subsample=0.8, nrof
## + Fold5: eta=0.20, max_depth=7, gamma=1, colsample_bytree=0.8, min_child_weight=1, subsample=1.0, nrof
## [12:02:25] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [12:02:25] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## - Fold5: eta=0.20, max_depth=7, gamma=1, colsample_bytree=0.8, min_child_weight=1, subsample=1.0, nrof
## + Fold5: eta=0.20, max_depth=7, gamma=1, colsample_bytree=1.0, min_child_weight=1, subsample=0.8, nrof
## [12:02:26] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [12:02:26] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## - Fold5: eta=0.20, max_depth=7, gamma=1, colsample_bytree=1.0, min_child_weight=1, subsample=0.8, nrof
## + Fold5: eta=0.20, max_depth=7, gamma=1, colsample_bytree=1.0, min_child_weight=1, subsample=1.0, nrof
## [12:02:27] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [12:02:27] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## - Fold5: eta=0.20, max_depth=7, gamma=1, colsample_bytree=1.0, min_child_weight=1, subsample=1.0, nrof
## + Fold5: eta=0.20, max_depth=7, gamma=5, colsample_bytree=0.8, min_child_weight=1, subsample=0.8, nrof
## [12:02:27] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [12:02:27] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## - Fold5: eta=0.20, max_depth=7, gamma=5, colsample_bytree=0.8, min_child_weight=1, subsample=0.8, nrof
## + Fold5: eta=0.20, max_depth=7, gamma=5, colsample_bytree=0.8, min_child_weight=1, subsample=1.0, nrof
## [12:02:28] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [12:02:28] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## - Fold5: eta=0.20, max_depth=7, gamma=5, colsample_bytree=0.8, min_child_weight=1, subsample=1.0, nrof
## + Fold5: eta=0.20, max_depth=7, gamma=5, colsample_bytree=1.0, min_child_weight=1, subsample=0.8, nrof
## [12:02:29] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [12:02:29] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## - Fold5: eta=0.20, max_depth=7, gamma=5, colsample_bytree=1.0, min_child_weight=1, subsample=1.0, nrof
## # Aggregating results
## Selecting tuning parameters
## Fitting nrounds = 100, max_depth = 7, eta = 0.01, gamma = 0, colsample_bytree = 0.8, min_child_weight = 1, nrof

```

```

# Best Parameters
cat("Best Parameters Found:\n")

```

```

## Best Parameters Found:

```

```

print(xgb_model$bestTune)

```

```

##      nrounds max_depth  eta gamma colsample_bytree min_child_weight subsample
## 74       100         7  0.01     0             0.8                 1        0.8

```

```

# Compute Precision, Recall, F1 manually using predictions
library(caret)
predictions <- xgb_model$pred

```

```

predictions <- predictions[predictions$Resample %in% paste0("Fold", 1:5), ]

# Confusion Matrix and Metrics
precision <- c()
recall <- c()
f1 <- c()
accuracy <- c()

for (i in 1:5) {
  fold_pred <- predictions[predictions$Resample == paste0("Fold", i), ]
  cm <- confusionMatrix(fold_pred$pred, fold_pred$obs, positive = "Yes")
  precision <- c(precision, cm$byClass["Precision"])
  recall <- c(recall, cm$byClass["Recall"])
  f1 <- c(f1, cm$byClass["F1"])
  accuracy <- c(accuracy, cm$overall["Accuracy"])
}

# Print Results
cat("\nXGBoost Results After Hyperparameter Tuning:\n")

```

```

##  
## XGBoost Results After Hyperparameter Tuning:  


```

```

cat(sprintf("Accuracy: %.4f\n", mean(accuracy)))

```

```

## Accuracy: 0.9473

```

```

cat(sprintf("Precision: %.4f\n", mean(precision)))

```

```

## Precision: 0.9541

```

```

cat(sprintf("Recall: %.4f\n", mean(recall)))

```

```

## Recall: 0.8418

```

```

cat(sprintf("F1-Score: %.4f\n", mean(f1)))

```

```

## F1-Score: 0.8944

```