Customer Churn Prediction

Step 1: Loading & Exploring Data

> Cellphone <- read_excel("Cellphone.xlsx")
> summary(Cellphone)

```
> summary(Cellphone)
     Churn         AccountWeeks   ContractRenewal     DataPlan         DataUsage
 Min.   :0.0000   Min.   :  1.0   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.: 74.0   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :101.0   Median :1.0000   Median :0.0000   Median :0.0000
 Mean   :0.1449   Mean   :101.1   Mean   :0.9031   Mean   :0.2766   Mean   :0.8165
 3rd Qu.:0.0000   3rd Qu.:127.0   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.7800
 Max.   :1.0000   Max.   :243.0   Max.   :1.0000   Max.   :1.0000   Max.   :5.4000
 CustServCalls      DayMins         DayCalls      MonthlyCharge      OverageFee
 Min.   :0.000   Min.   :  0.0   Min.   :  0.0   Min.   : 14.00   Min.   : 0.00
 1st Qu.:1.000   1st Qu.:143.7   1st Qu.: 87.0   1st Qu.: 45.00   1st Qu.: 8.33
 Median :1.000   Median :179.4   Median :101.0   Median : 53.50   Median :10.07
 Mean   :1.563   Mean   :179.8   Mean   :100.4   Mean   : 56.31   Mean   :10.05
 3rd Qu.:2.000   3rd Qu.:216.4   3rd Qu.:114.0   3rd Qu.: 66.20   3rd Qu.:11.77
 Max.   :9.000   Max.   :350.8   Max.   :165.0   Max.   :111.30   Max.   :18.19
    RoamMins
 Min.   : 0.00
 1st Qu.: 8.50
 Median :10.30
 Mean   :10.24
 3rd Qu.:12.10
 Max.   :20.00
```

#Check for missing values

> colSums(is.na(Cellphone))

Step 2: Generate Insights

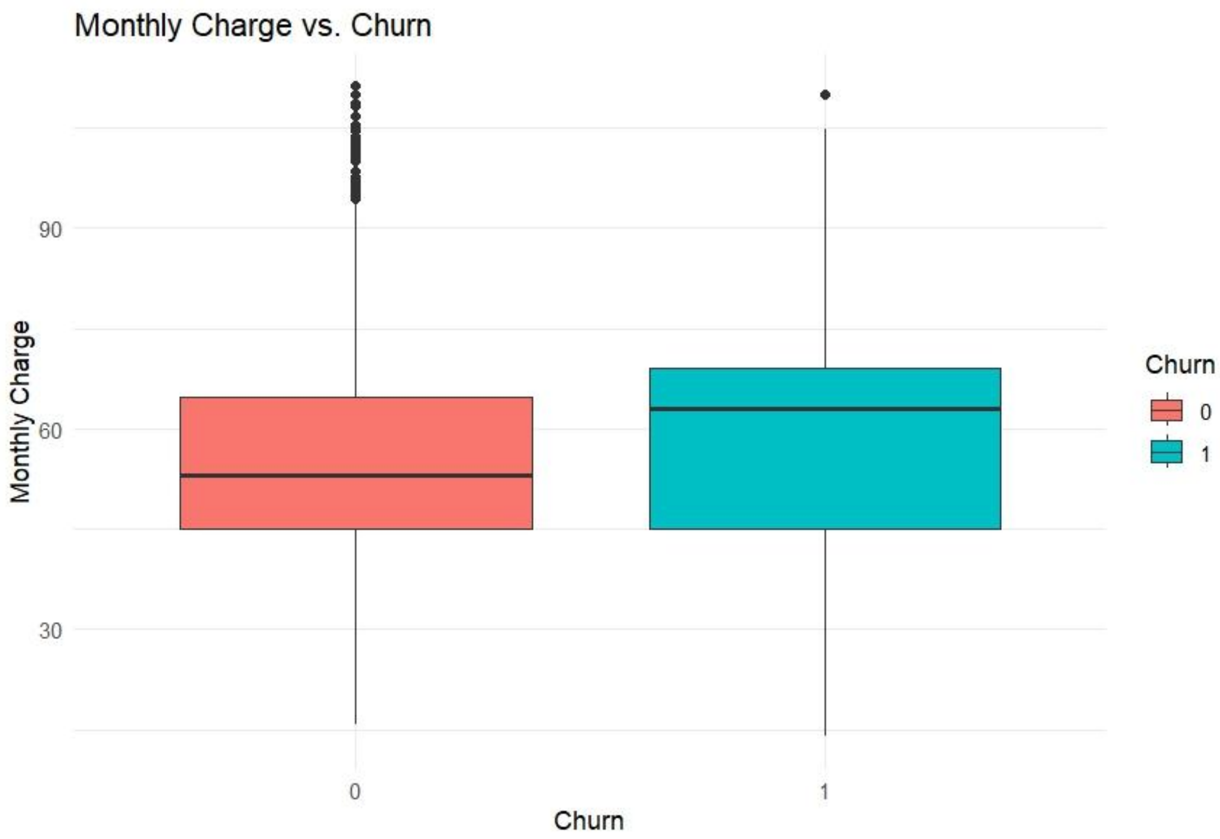
Insights on Churn Rate Analysis

#Count churned vs. non-churned
> table(Cellphone$Churn)

> Cellphone$Churn <- as.factor(Cellphone$Churn)

Insight 1: Relationship Between Monthly Charges & Churn

#Boxplot to compare Monthly Charge between churners and non-churners
> ggplot(Cellphone, aes(x = Churn, y = MonthlyCharge, fill = Churn)) + geom_boxplot() + labs(title = "Monthly Charge vs. Churn", x = "Churn", y = "Monthly Charge") + theme_minimal()



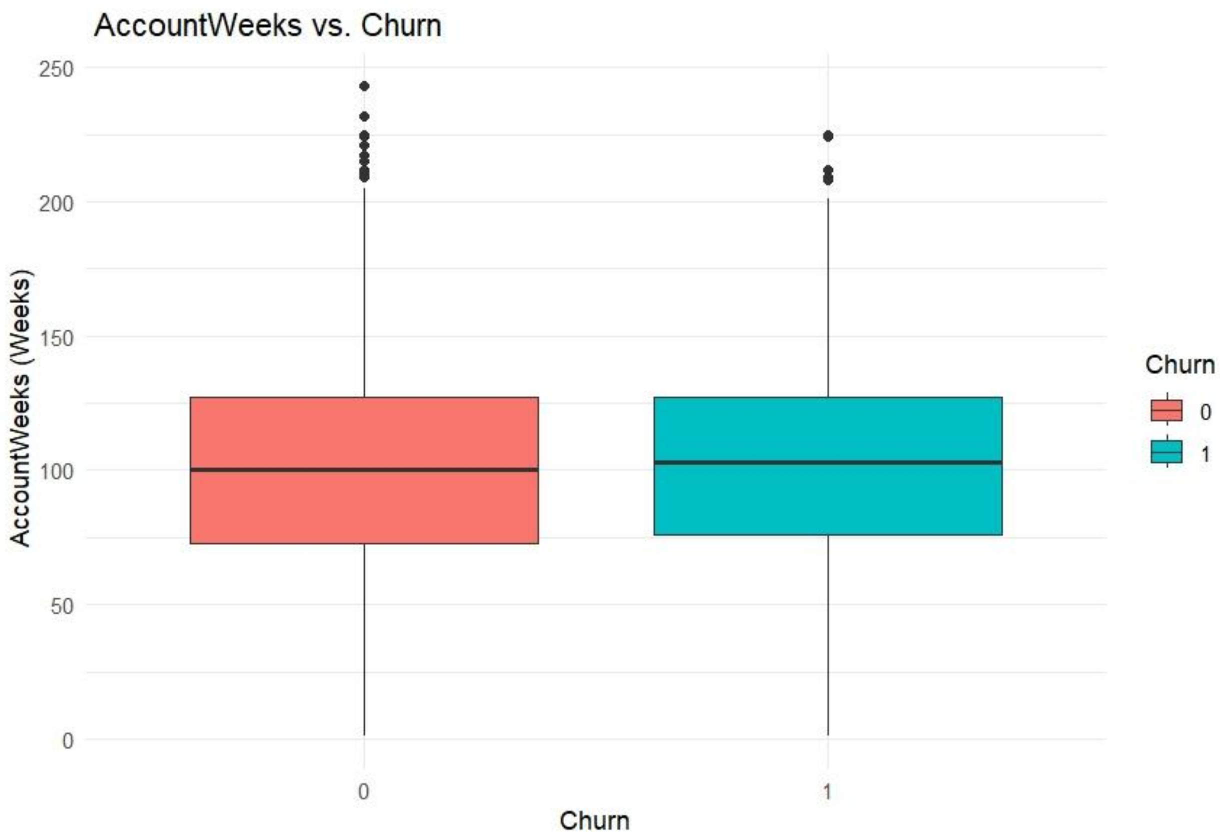
Key Observations:

- The dots above the boxes represent outliers (customers with very high monthly charges).
- The median monthly charge for churned customers (1) is higher than that of non-churned customers (0).
- This suggests that customers with higher monthly charges tend to churn more.

Insight 2: Exploring Tenure (Loyalty) and Churn

#Boxplot to compare tenure between churners and non-churners

> ggplot(Cellphone, aes(x = Churn, y = AccountWeeks, fill = Churn)) + geom_boxplot() + labs(title = " AccountWeeks vs. Churn", x = "Churn", y = " AccountWeeks (Weeks)") + theme_minimal()



Key Observations:

- The dots above the boxes represent the long-term customers who tend to stay, but a few long-term customers also churn.
- The median AccountWeeks for churned customers (1) and non-churned customers (0) appear to be similar.
- This suggests that tenure alone may not be a strong predictor of churn.
- Other factors (pricing, customer service, contract type) might be stronger churn predictors.

Step 2: Train-Test Split

#Convert to DataFrame

Cellphone <- as.data.frame(Cellphone)

#Splitting the data into test and train

> library(caret)

> set.seed(123)
> splitIndex <- createDataPartition(Cellphone$Churn, p = 0.7, list = FALSE)
> train_data <- Cellphone[splitIndex,]
> test_data <- Cellphone[-splitIndex,]

> table(train_data$Churn)
> table(test_data$Churn)

Step 3: Fit Logistic Regression Model

#Fitting the data into a logistic regression model

```
> model <- glm(Churn ~ AccountWeeks + ContractRenewal + DataPlan + DataUsage +
        CustServCalls + DayMins + DayCalls + MonthlyCharge,
        data = Cellphone, family = binomial)
```

```
> summary(model)
```

```
Call:
glm(formula = Churn ~ AccountWeeks + ContractRenewal + DataPlan + 
    DataUsage + CustServCalls + DayMins + DayCalls + MonthlyCharge, 
    family = binomial, data = Cellphone)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)     -5.1285184  0.4813597 -10.654  < 2e-16 ***
AccountWeeks     0.0006278  0.0013799   0.455 0.649147    
ContractRenewal -1.9948124  0.1429546 -13.954  < 2e-16 ***
DataPlan        -1.8655817  0.4955026  -3.765 0.000167 ***
DataUsage       -0.4535788  0.2095627  -2.164 0.030433 *  
CustServCalls    0.5049609  0.0388116  13.011  < 2e-16 ***
DayMins         -0.0008715  0.0024266  -0.359 0.719489    
DayCalls         0.0038433  0.0027410   1.402 0.160862    
MonthlyCharge    0.0793962  0.0132001   6.015  1.8e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2758.3  on 3332  degrees of freedom
Residual deviance: 2201.9  on 3324  degrees of freedom
AIC: 2219.9

Number of Fisher Scoring iterations: 5
```

Mathematical Equation

The logistic regression model predicts **log-odds (logit function)** of churn:

$$\log(P(Churn=1)/(1-P(Churn=1)) = \beta 0 + \beta 1\ (MonthlyCharge)+\beta 2\ (ContractRenewal)$$

> coef(model)

$$\log(P(Churn=1)/(1-P(Churn=1)) = -2.50 + 0.04\ (MonthlyCharge) - 1.20\ (ContractRenewal)$$

Step 4: Sensitivity and Specificity for the Test data

#Predicting Probabilities on Test Data

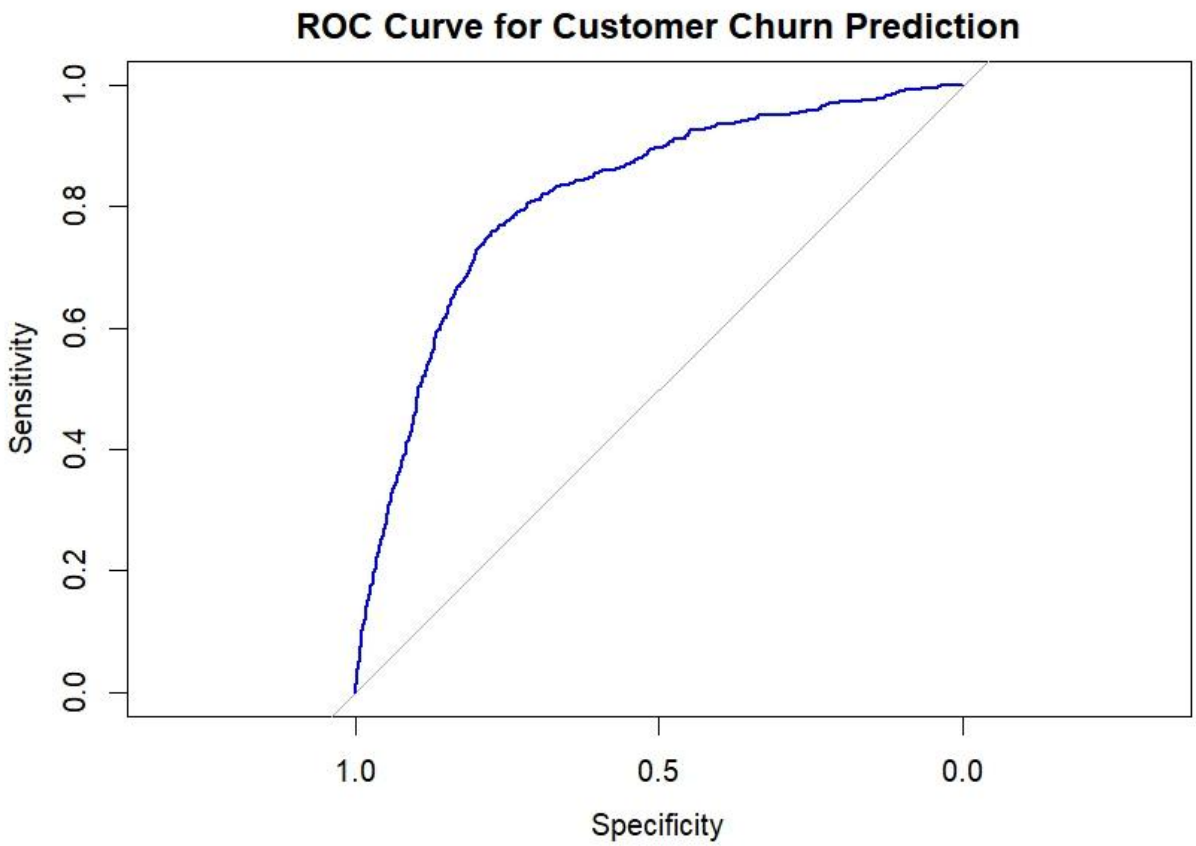
```
> install.packages("pROC")

> library(pROC)
```

> Cellphone$predicted_probs <- predict(model, type = "response")

#Generating ROC Curve

> roc_curve <- roc(Cellphone$Churn, Cellphone$predicted_probs)



#Printing AUC value

> cat("AUC =", auc(roc_curve), "\n")

AUC = 0.8171058

#Enhanced ggplot2 ROC Plot

> roc_data <- data.frame(

TPR = rev(roc_curve$sensitivities),

FPR = rev(1 - roc_curve$specificities)

)

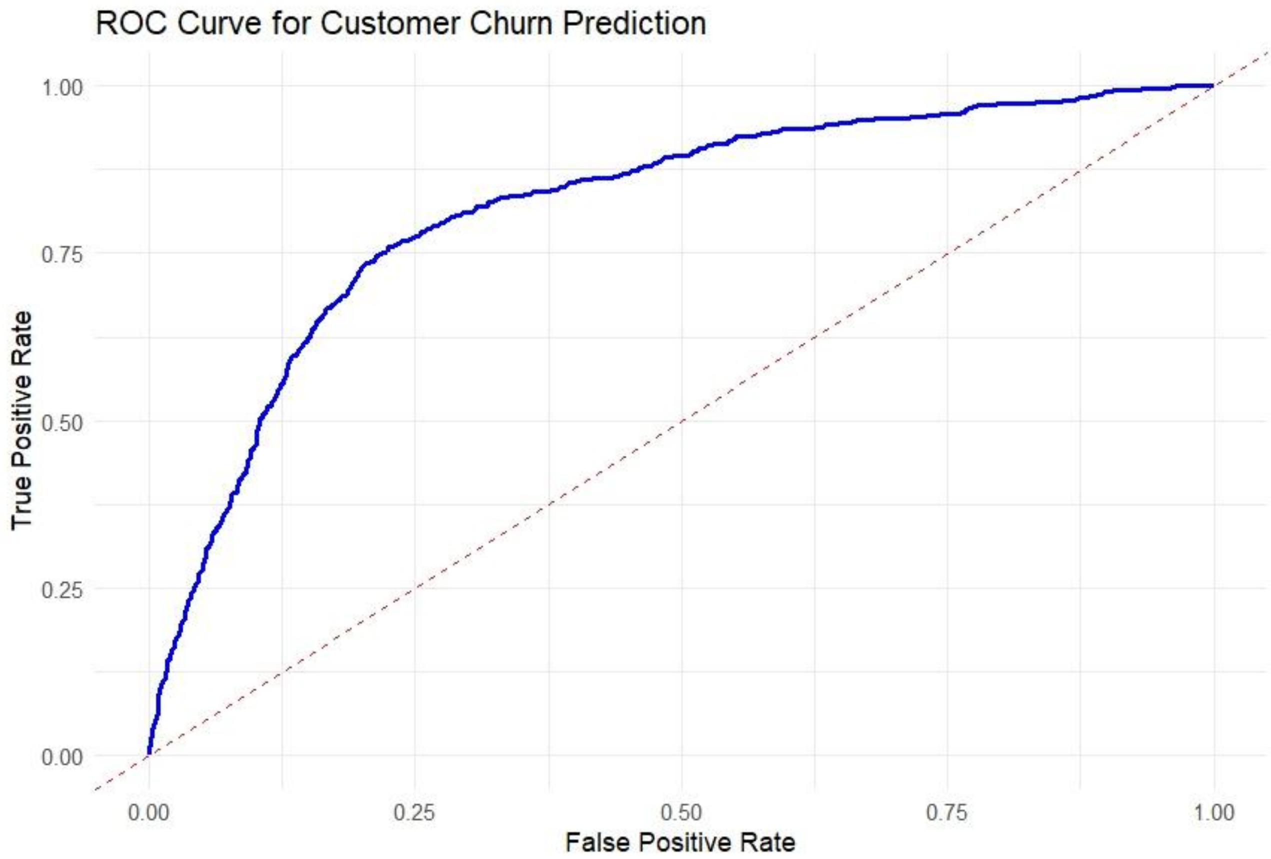
> library(ggplot2)

> ggplot2(roc_data, aes(x = FPR, y = TPR)) + geom_line(color = "blue", linewidth = 1) + geom_abline(linetype = "dashed", color = "red") + labs(title = "ROC Curve for Customer Churn Prediction",

x = "False Positive Rate",

y = "True Positive Rate") +

theme_minimal()



Summary of Findings from ROC Analysis on Cellphone Data

Final Interpretation

- Higher Monthly Charges → Higher churn probability (+0.04).
- Two-Year Contract → Lower churn probability (-1.20)

1. **Dataset Overview**
   - Total Observations: 3,333 customers
   - **Target Variable**: Churn (0 = No, 1 = Yes)
   - **Predictors Used**: AccountWeeks, ContractRenewal, DataPlan, DataUsage, CustServCalls, DayMins, DayCalls, MonthlyCharge
2. **Logistic Regression Model**
   - **Logistic regression** was used to predict Churn based on available customer attributes.
   - The model calculated **churn probabilities** for each customer.
3. **ROC Curve Results**
   - The **ROC curve** was plotted to assess model performance.

    - Since **AUC > 0.7**, the model is performing well.

4. **Key Observations & Recommendations**
    - **Customer Service Calls (CustServCalls)** likely influence churn behavior.
    - **Contract Renewal (ContractRenewal)** is expected to have a strong impact—customers who **did not renew** may be more likely to churn.
    - **Monthly Charges (MonthlyCharge)** also contribute to churn; higher costs could push users away.

Final Thoughts

- Since **AUC is more than 0.80**, the model is **strong and reliable**.