MACT6100 Assessment 2

Junchi Han

2025-12-04

1. Introduction

1.1 Introduction to Customer Churn

Customer churn is a major challenge in the telecommunications industry due to intense competition and low switching costs. Existing research shows that churn rates are directly related to the revenue.

It is also suggested that retaining customers is more cost-effective than acquiring new ones. As a result, churn prediction has become a key application of machine learning and statistical modelling in both academia and industry.

1.2 Background of Customer Churn Data

Previous academic studies has demonstrated that classification methods can effectively predict customer churn using behavioural and demographic data.

This study uses the Telco Customer Churn dataset from Kaggle, which contains rich information on customer demographics, service usage, contracts, and billing behaviour. The topic was chosen due to its strong real-world relevance and its suitability for comparing machine-learning techniques in a practical setting.

1.3 Aim of the Research

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

1.4 Methology & Models

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

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

The models considered include the traditional statistical approach of GLM, as well as machine learning and deep learning models such as Decision Tree, Neural Network, and Random Forest.

All models are evaluated and compared using performance metrics including ROC, AUC and accuracy, to provide a comprehensive assessment of their effectiveness in customer churn prediction.

2. Data Loading & Exploring

2.1 Load Customer Churn Data

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

```r
library(tidyverse)

library(dplyr)

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

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

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

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

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

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

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

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

Noticed that 11 N/A values have been removed.

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

```
sdata <- sdata %>% dplyr::select(-customerID)
```
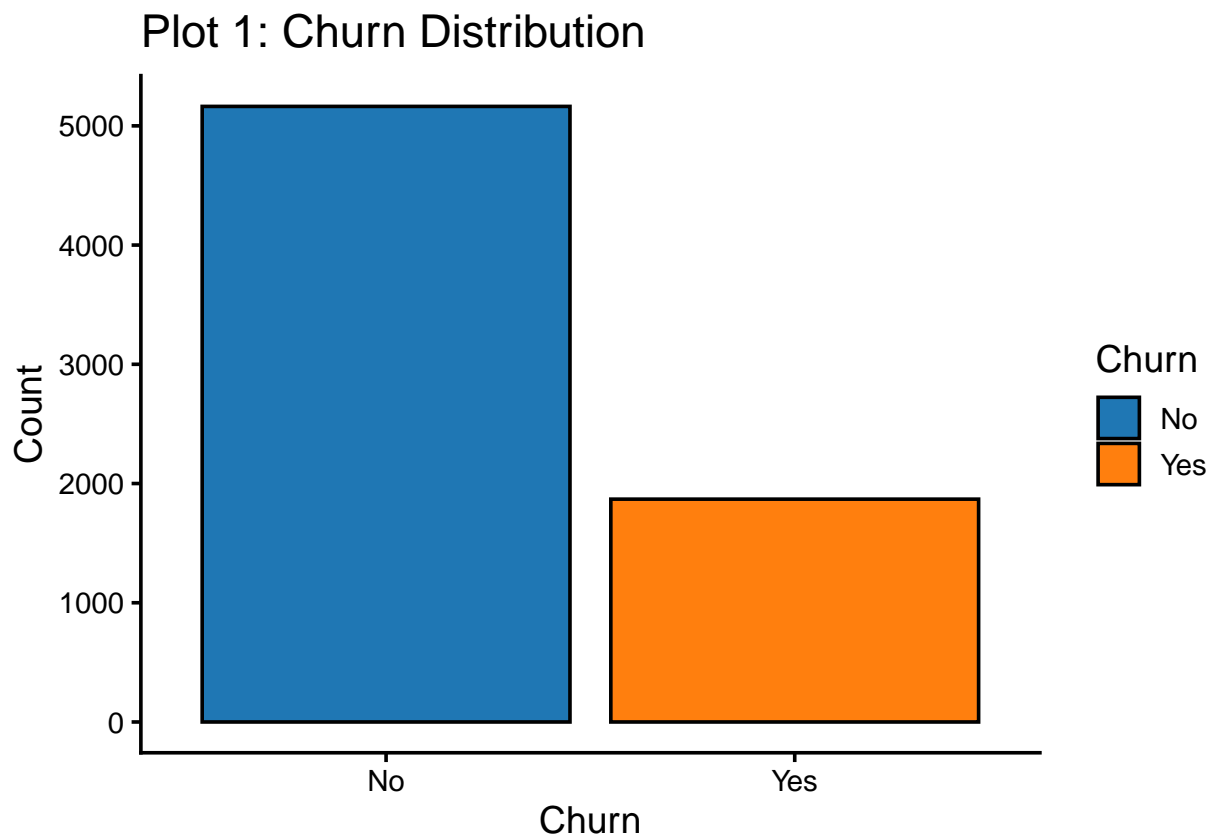
2.2 Exploratory Data Analysis on Churn Data

After initial data explore, we want to have understanding of how the elements can be related to each other. Graph is a good method to help to build understanding by visualising the correlations.
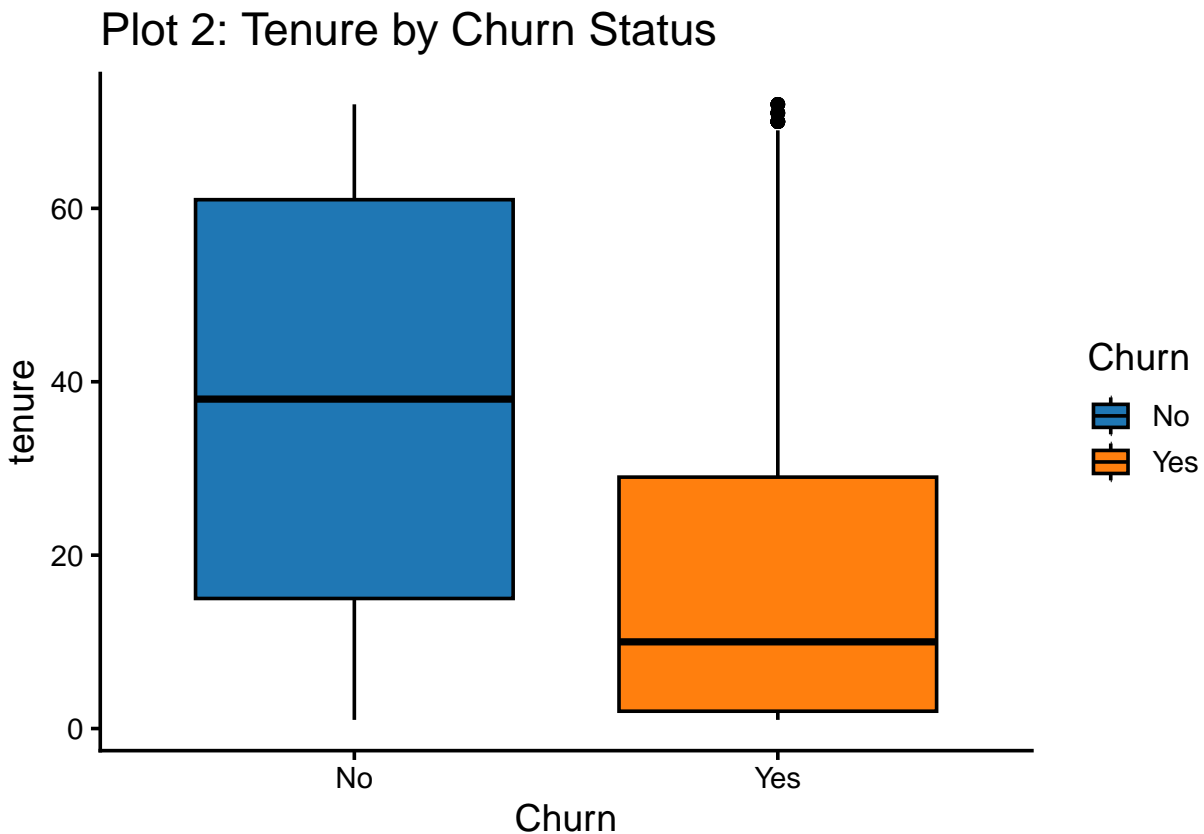
```
library(ggplot2)

library(corrplot)

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


Plot 1:

- The amount of customer classified as "non-churned" is almost triple than those classified as "churn".
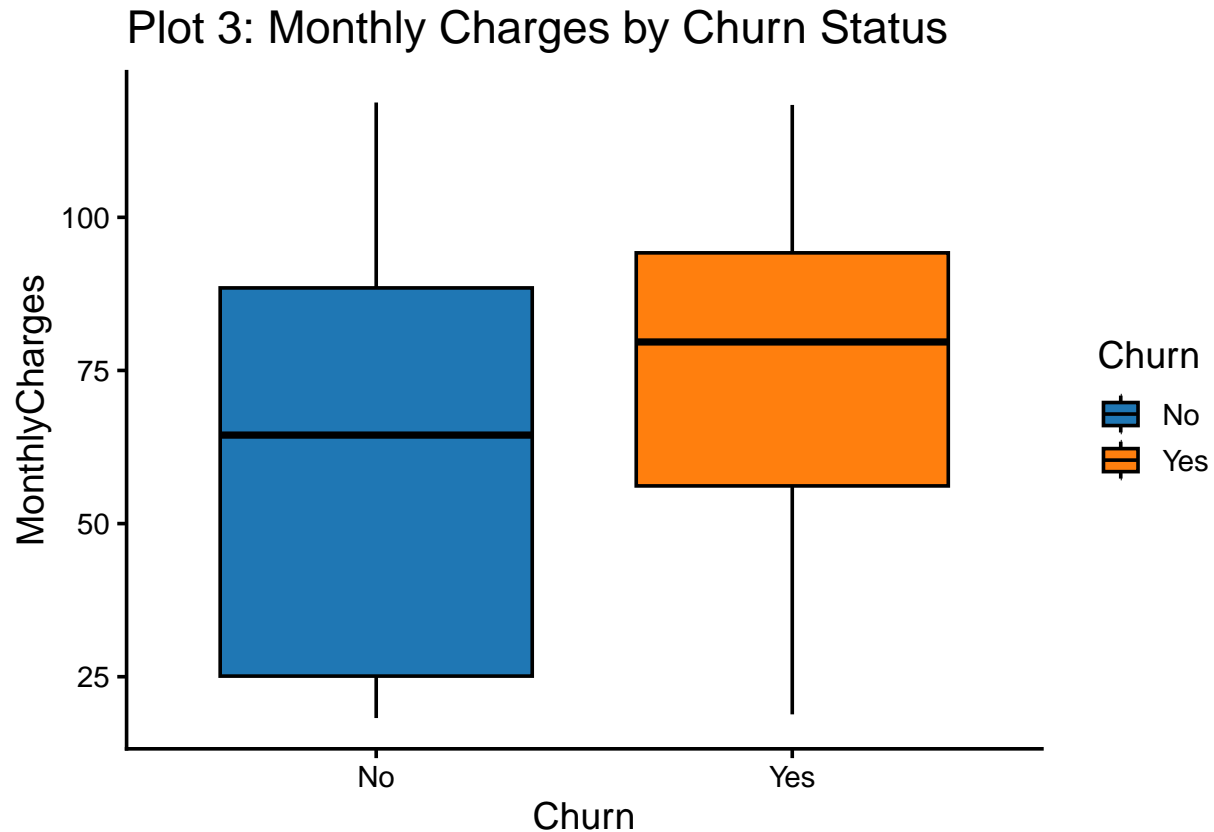
```
ggplot(sdata, aes(Churn, tenure, fill = Churn)) +
  geom_boxplot(color="black") +
  scale_fill_manual(values=c("No"="#1f77b4", "Yes"="#ff7f0e")) +
  labs(title = "Plot 2: Tenure by Churn Status") +
  theme_classic(base_size = 14)
```


Plot 2:

- The tenure distribution of churned customers (Yes) is clearly lower, with most concentrated among short-term users (0–20 months). In contrast, non-churned customers generally have much longer usage periods, and long-term customers have a lower risk of churn.

- This indicates that the shorter the usage period, the more likely customers are to churn, reflecting the characteristics of "new users being less stable and having lower loyalty".
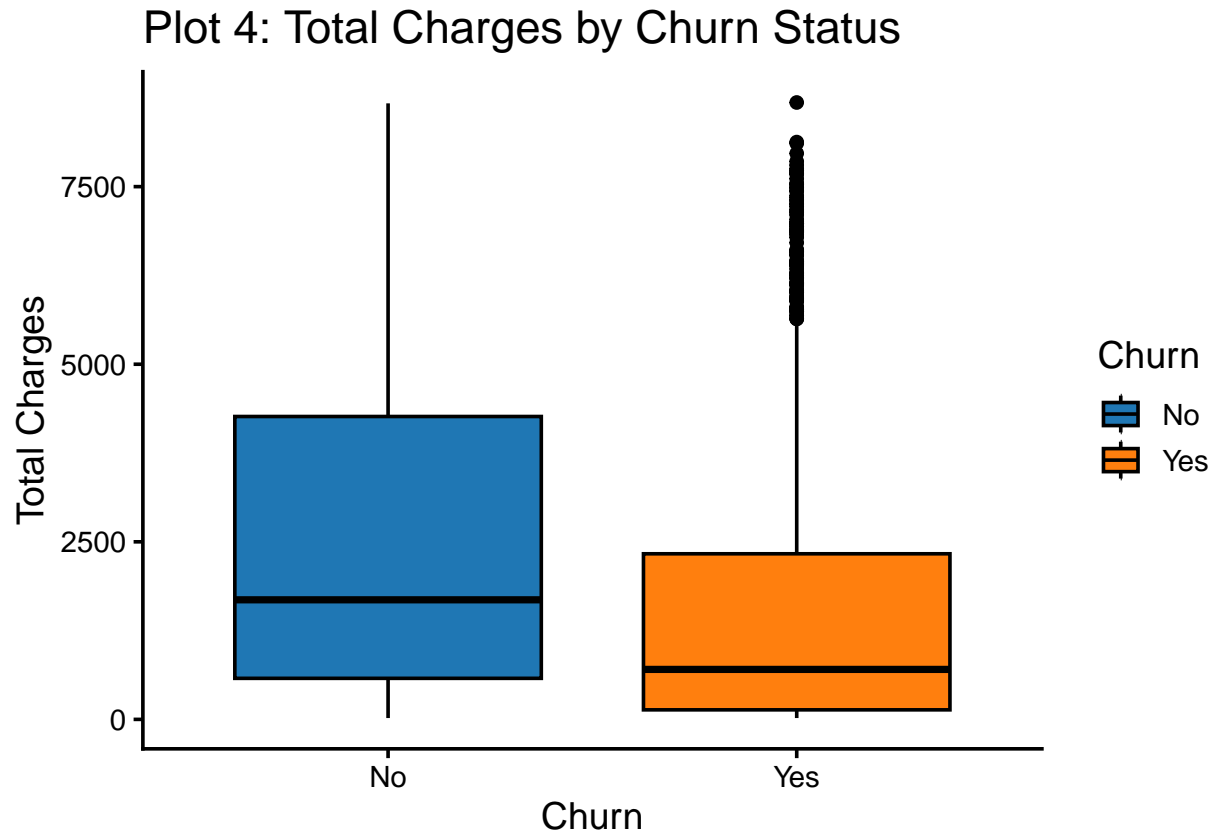
```
ggplot(sdata, aes(Churn, MonthlyCharges, fill = Churn)) +
  geom_boxplot(color="black") +
  scale_fill_manual(values=c("No"="#1f77b4", "Yes"="#ff7f0e")) +
  labs(title = "Plot 3: Monthly Charges by Churn Status") +
  theme_classic(base_size = 14)
```

Plot 3: Monthly Charges by Churn Status

Plot 3:

- Customers who churn tend to have higher monthly charges than those who remain, as shown by the higher median for the churned group.

- Although there is some overlap between the two groups, the overall distribution is shifted upward for churned customers, suggesting that higher prices are associated with a greater likelihood of churn.

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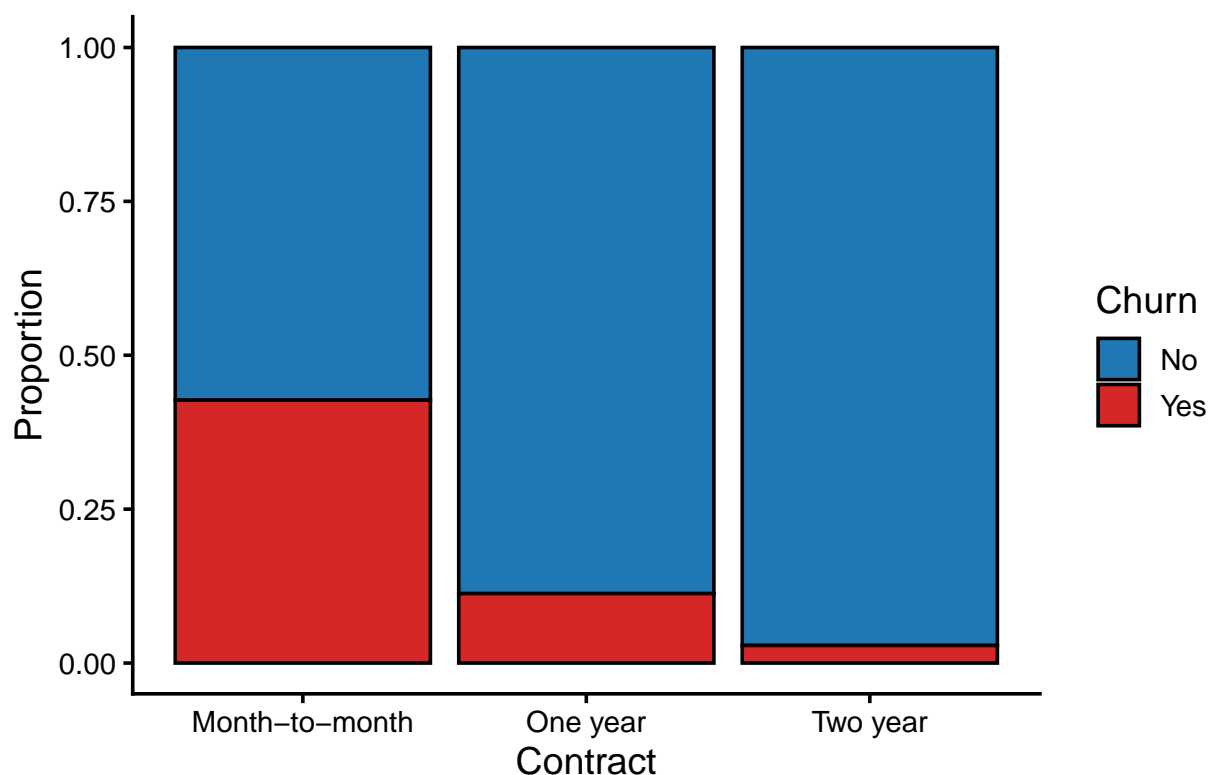
Plot 4: Total Charges by Churn Status

Plot 4:

- Customers who did not churn have much higher total charges on average, reflecting longer tenure, while churned customers generally show low total charges, indicating early-stage departure.

- However, the churned group contains several high-value outliers, showing that although rare, some long-standing, high-spending customers still choose to leave, possibly due to dissatisfaction or external competition.

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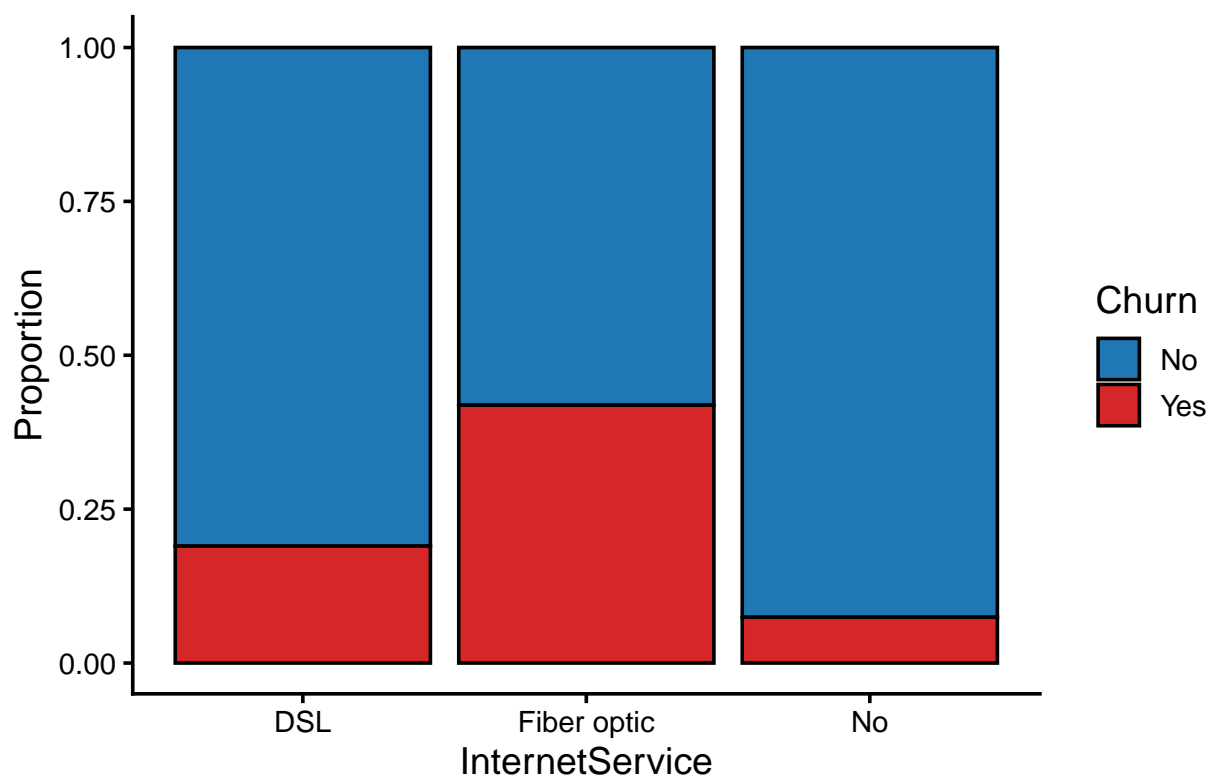
Plot 5: Churn Rate by Contract Type

Plot 5:

- Customers on month-to-month contracts exhibit the highest churn rate, while those on one-year and especially two-year contracts are far more likely to remain.

- This indicates that longer contract durations are strongly associated with improved customer retention, highlighting contract length as a key driver of churn.

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

Plot 6: Churn Rate by Internet Service Type

Plot 6:

- Customers with fiber optic internet have the highest churn rate, while DSL users show moderate churn, and customers with no internet service have the lowest churn.

- This suggests that service type is strongly associated with churn, with fiber optic customers being particularly at risk.

```
library(reshape2)

numdata <- sdata

numdata$TotalCharges <- as.numeric(numdata$TotalCharges)

num_vars <- numdata[, c("tenure", "MonthlyCharges", "TotalCharges")]

cor_mat <- cor(num_vars, use = "complete.obs")
cor_mat
```
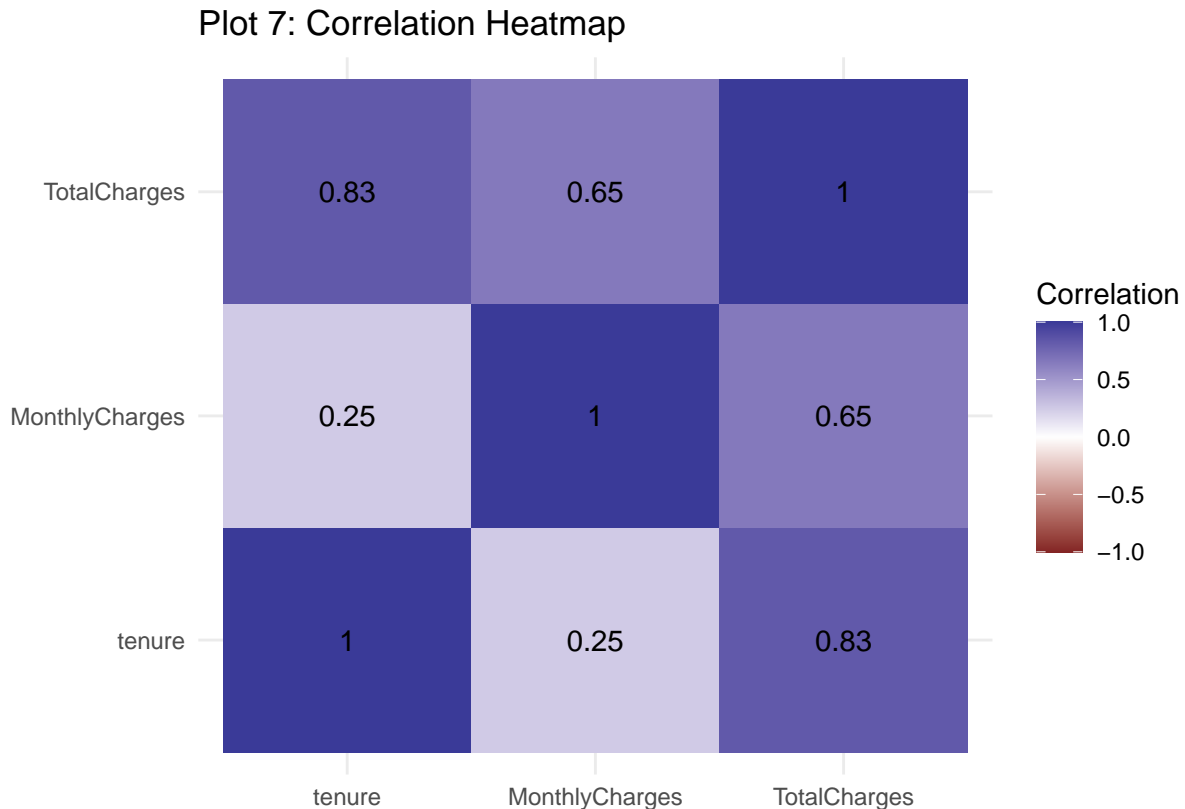
```
##                  tenure MonthlyCharges TotalCharges
## tenure         1.0000000      0.2468618    0.8258805
## MonthlyCharges 0.2468618      1.0000000    0.6510648
## TotalCharges   0.8258805      0.6510648    1.0000000
```

```
cor_df <- melt(cor_mat)

ggplot(cor_df, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  geom_text(aes(label = round(value, 2))) +
  scale_fill_gradient2(limits = c(-1, 1)) +
  labs(title = "Plot 7: Correlation Heatmap",
       x = "", y = "", fill = "Correlation") +
  theme_minimal()
```

Plot 7: Correlation Heatmap

Plot 7:

- The heatmap shows a strong positive correlation between tenure and total charges (0.83), indicating that customers who stay longer naturally accumulate higher total charges.

- Monthly charges are moderately correlated with total charges (0.65) but only weakly correlated with tenure (0.25), suggesting that how long a customer stays matters more for total spending than the monthly price alone.

3. Model Fitting and Analysis

3.1 Data Preparation for Model Training

From the previous results we learned that the variables are stored in different forms. In this case, we want the target variable "Churn" to be converted into a factor to ensure that the subsequent classification models

correctly recognise its categorical nature.

We then use the createDataPartition function from the caret package to split the dataset into a training set and a test set in an 8:2 ratio, where the training set is used for model fitting and the test set for independent performance evaluation.

By setting a random seed (set.seed), the reproducibility of the data split is ensured, so that the same partition is obtained across different runs. This process establishes a consistent and reliable basis for the training and evaluation of all subsequent models.

```r
library(caret)


sdata$Churn <- as.factor(sdata$Churn)

### Train-Test split (8:2)
set.seed(183)
train_index <- createDataPartition(sdata$Churn, p = 0.8, list = FALSE)
train <- sdata[train_index, ]
test <- sdata[-train_index, ]
```

We also want a function that can produce ROC curve for analitical purposes:

```r
library(pROC)

plot_roc_with_auc <- function(model_name, prob_train, train_labels,
                              prob_test, test_labels) {

  ### Train AUC
  roc_train <- roc(train_labels, prob_train)
  auc_train <- auc(roc_train)

  ### Test AUC
  roc_test <- roc(test_labels, prob_test)
  auc_test <- auc(roc_test)

  ### Plot
  plot(roc_train,
       col="#1f77b4", lwd=2,
       main=paste(model_name, "ROC Curve"),
       legacy.axes = TRUE)
  lines(roc_test, col="#ff7f0e", lwd=2)

  legend("bottomright",
         legend=c(
           paste("Train AUC =", round(auc_train, 3)),
           paste("Test AUC  =", round(auc_test, 3))
         ),
         col=c("#1f77b4", "#ff7f0e"),
         lwd=2)
}
```
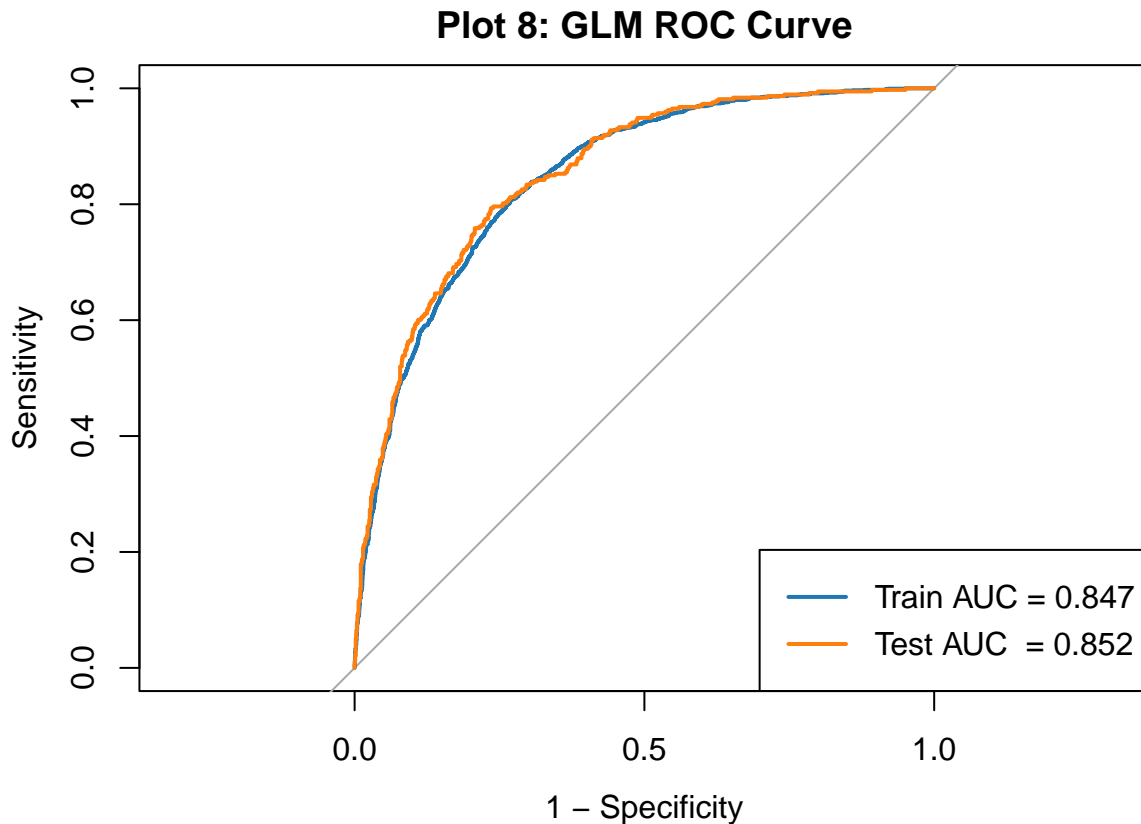
3.2 GLM: Logistic Regression Model Fitting

In this section, a generalised linear model (GLM) with a logistic link function is used to model customer churn because the response variable is binary (churn vs non-churn).

Logistic regression is well-suited for this setting as it estimates the probability of churn while allowing the effect of multiple explanatory variables to be quantified and interpreted. It is chosen for its simplicity and interpretability, making it an ideal baseline model for churn prediction.

```r
### Train GLM
set.seed(183)
glm_model <- glm(Churn ~ ., data=train, family=binomial)

### Predictions
glm_train_prob <- predict(glm_model, train, type="response")
glm_test_prob  <- predict(glm_model, test,  type="response")

### ROC
plot_roc_with_auc("Plot 8: GLM",
                  glm_train_prob, train$Churn,
                  glm_test_prob, test$Churn)
```


Plot 8:

- The logistic GLM shows good discriminative performance, with a training AUC of 0.847 and a test AUC of 0.852.

11

- The close agreement between training and test AUC indicates strong generalisation and no obvious evidence of overfitting, suggesting that the model is reliable for predicting customer churn.

- It is sensible to use GLM model as baseline model for churn prediction.

3.3 Decision Tree Model Fitting

In this section, a decision tree model is fitted as a non-parametric classification method that captures non-linear relationships and variable interactions in the churn data. It is chosen for its interpretability, ability to handle mixed data types, and for providing a useful comparison to the parametric logistic GLM, allowing the performance of linear and non-linear models to be compared directly.
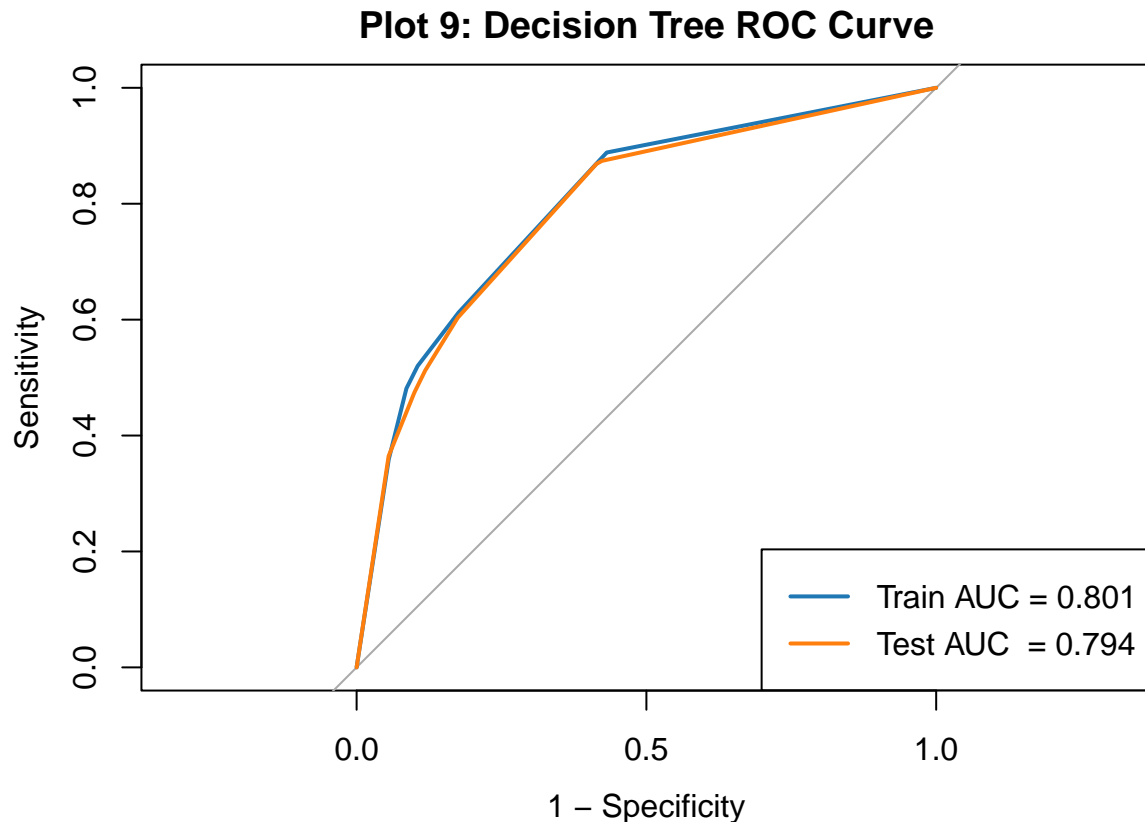
```r
library(rpart)

set.seed(183)

tree_model <- rpart(Churn ~ ., data=train, method="class")

### Train-test split
tree_train_prob <- predict(tree_model, train, type="prob")[,2]

tree_test_prob  <- predict(tree_model, test,  type="prob")[,2]


plot_roc_with_auc("Plot 9: Decision Tree",
                  tree_train_prob, train$Churn,
                  tree_test_prob,  test$Churn)
```

Plot 9: Decision Tree ROC Curve

Plot 9:

- The decision tree model shows good predictive performance, with a training AUC of 0.801 and test AUC of 0.794.

- The close values indicate good generalisation with no obvious overfitting, although its overall discrimination ability is weaker than the logistic GLM, suggesting the tree captures non-linear patterns but with reduced predictive strength.

3.4 Neural Network Model Fitting

In this section, a neural network model is fitted to capture complex non-linear relationships and interactions among the customer features that simpler models may not fully represent. It is chosen for its flexibility and strong predictive capability, providing a powerful contrast to traditional models such as logistic regression and decision trees for churn prediction.

```
library(nnet)

library(caret)

### Introduce dummy variables
dmy <- dummyVars(" ~ .", data=train)


### Train-test split
train_nn <- data.frame(predict(dmy, newdata=train))
```

```r
test_nn  <- data.frame(predict(dmy, newdata=test))


### Standardise the scale of the input variables, to make the model less sensitive
pre <- preProcess(train_nn, method=c("center", "scale"))

train_nn <- predict(pre, train_nn)

test_nn  <- predict(pre, test_nn)

### Remove Churn.No generated by dummyVars
train_nn = train_nn %>% dplyr::select(-Churn.No)

test_nn  = test_nn  %>% dplyr::select(-Churn.No)


### Yes → 1, No → 0
train_nn$Churn.Yes <- ifelse(train$Churn == "Yes", 1, 0)

test_nn$Churn.Yes  <- ifelse(test$Churn == "Yes", 1, 0)

### Train neuro network
set.seed(183)
nn_model <- nnet(
  Churn.Yes ~ .,
  data=train_nn,
  size=5,
  maxit=300,
  linout=FALSE,
  decay=1e-4,
  trace=FALSE
)


### Predict the churn rate
nn_train_prob <- predict(nn_model, train_nn)
nn_test_prob  <- predict(nn_model, test_nn)

### Plot output
plot_roc_with_auc("Plot 10: Neural Network",
                  nn_train_prob, train_nn$Churn.Yes,
                  nn_test_prob,  test_nn$Churn.Yes)
```
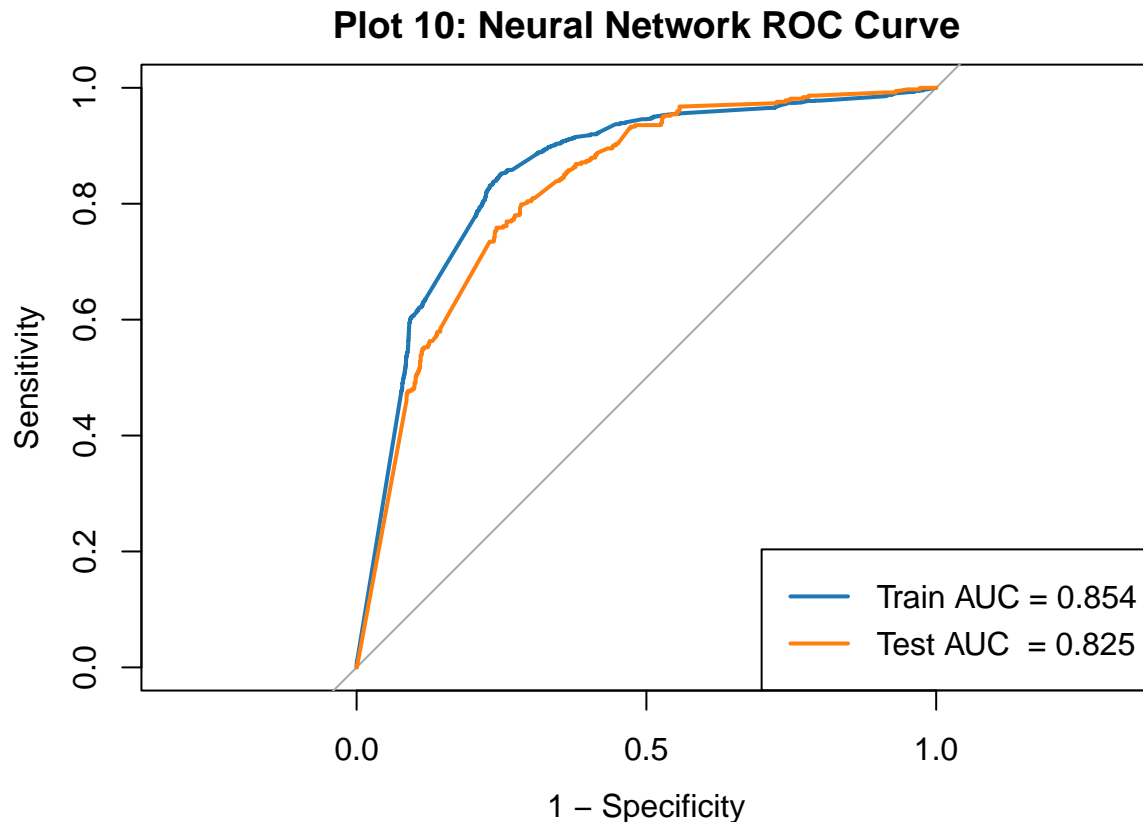
Plot 10: Neural Network ROC Curve

Legend:
- Train AUC = 0.854
- Test AUC = 0.825

Plot 10:

- The neural network achieves strong predictive performance with a training AUC of 0.854 and test AUC of 0.825, indicating good discrimination ability.

- The small drop from training to test AUC suggests mild overfitting, but overall the model performs well in capturing complex non-linear patterns in customer churn.

3.5 Random Forest Model Fitting

In this section, a random forest model is fitted as an alternative method that combines multiple decision trees to improve predictive accuracy and robustness. It is chosen for its ability to capture complex non-linear patterns and provide measures of variable importance for understanding key drivers of churn.
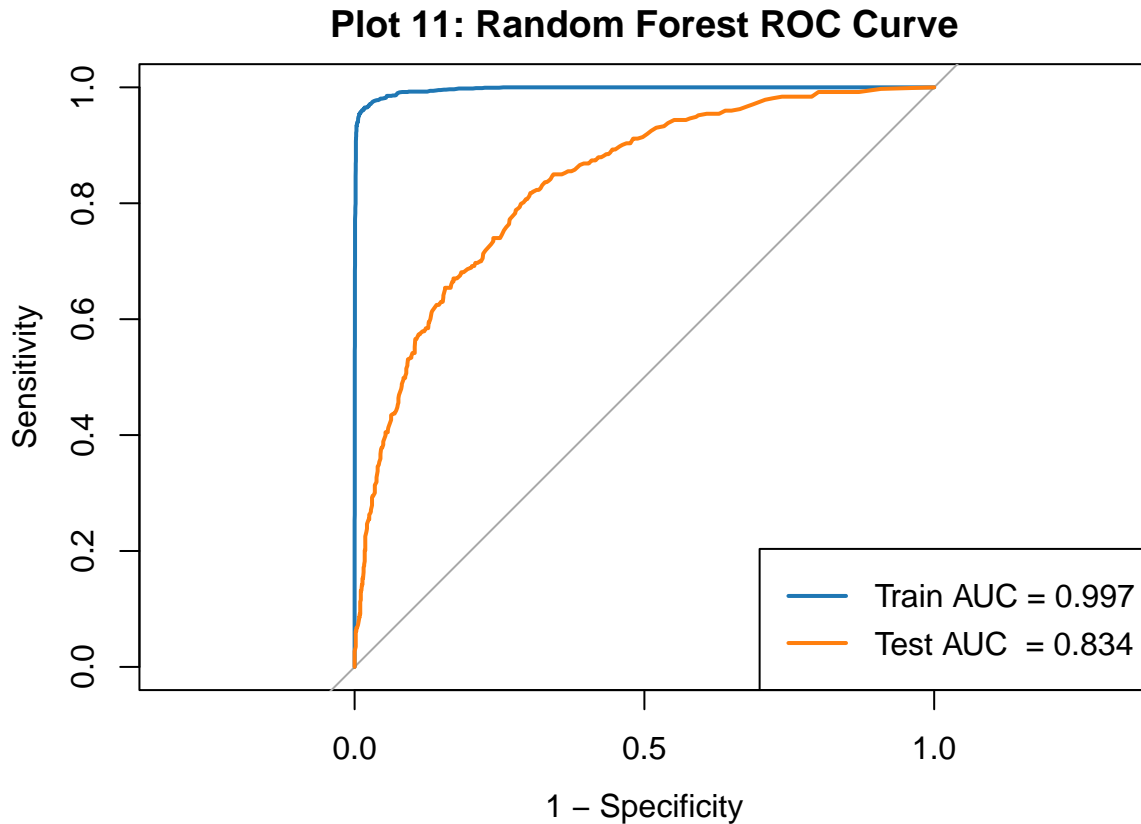
```r
library(randomForest)

set.seed(183)

rf_model <- randomForest(Churn ~ ., data=train, ntree=300)

### predict churn rate
rf_train_prob <- predict(rf_model, train, type="prob")[,2]

rf_test_prob  <- predict(rf_model, test,  type="prob")[,2]
```

```
plot_roc_with_auc("Plot 11: Random Forest",
                  rf_train_prob, train$Churn,
                  rf_test_prob, test$Churn)
```

Plot 11: Random Forest ROC Curve

Plot 10:

- The random forest achieves very high training performance (AUC = 0.997) and strong test performance (AUC = 0.834), indicating excellent fitting ability.
- However, the large gap between training and test AUC suggests overfitting, meaning the model captures training patterns extremely well but generalises less effectively to new data.

4. Model Comparison and Conclusion

4.1 Model Accuracy Comparison

In the previous section, we have used different models to predict the customer churn data. In this section, we will compare the model accuracy and suggest company's operational strategy based on our analysis.

```
# GLM accuracy
glm_pred <- ifelse(glm_test_prob > 0.5, "Yes", "No") %>% factor(levels=c("No","Yes"))

glm_cm <- confusionMatrix(glm_pred, test$Churn)
```

```r
glm_acc <- glm_cm$overall["Accuracy"]

# Decision Tree
tree_pred <- ifelse(tree_test_prob > 0.5, "Yes", "No") %>% factor(levels=c("No","Yes"))

tree_cm <- confusionMatrix(tree_pred, test$Churn)

tree_acc <- tree_cm$overall["Accuracy"]

# Neural Network
nn_pred <- ifelse(nn_test_prob > 0.5, 1, 0)

nn_cm <- confusionMatrix(
  factor(nn_pred, levels=c(0,1)),
  factor(test_nn$Churn.Yes, levels=c(0,1))
)

nn_acc <- nn_cm$overall["Accuracy"]

# Random Forest
rf_pred <- ifelse(rf_test_prob > 0.5, "Yes", "No") %>% factor(levels=c("No","Yes"))

rf_cm <- confusionMatrix(rf_pred, test$Churn)

rf_acc <- rf_cm$overall["Accuracy"]



acc <- data.frame(
  Model = c("GLM", "Decision Tree", "Neural Net", "Random Forest"),
  Accuracy = c(glm_acc, tree_acc, nn_acc, rf_acc)
)

### Plot output
ggplot(acc, aes(x = Model, y = Accuracy, fill = Model)) +
  geom_col(color="black") +
  geom_text(aes(label = sprintf("%.3f", Accuracy)),
            vjust = -0.5,
            size = 5) +
  scale_fill_manual(values=c("#1f77b4","#ff7f0e","#2ca02c","#d62728")) +
  labs(title = "Plot 11: Model Accuracy (Based on Confusion Matrix)",
       y="Accuracy") +
  theme_classic(base_size=14) +
  ylim(0, 1.1)
```
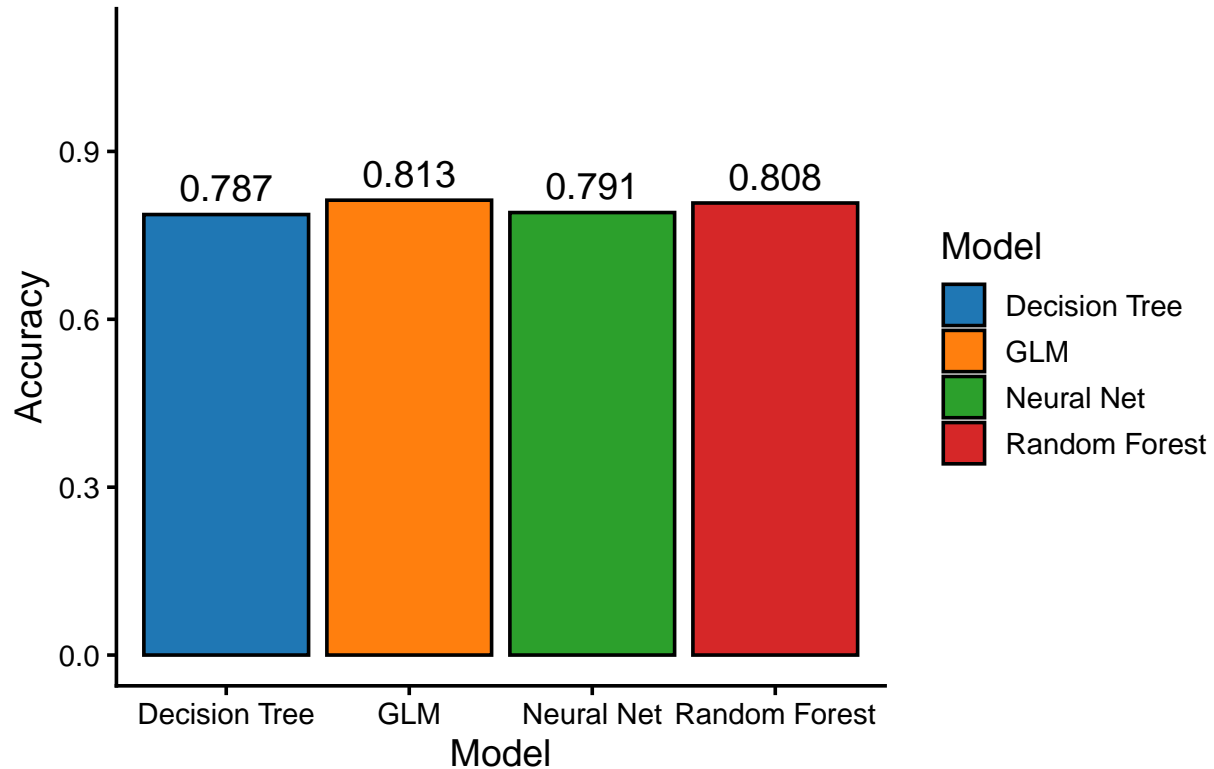
Plot 11: Model Accuracy (Based on Confusion Matrix)

Plot 11:

- The GLM achieves the highest accuracy (0.813), closely followed by the random forest (0.808), while the neural network (0.791) and decision tree (0.787) perform slightly worse.

- Overall, all models show similar accuracy, with GLM providing the best balance between performance and simplicity.

4.2 Important Features to Investigate

We also want to investigate what variables have the strongest effect on model's decision making. We use random forest model as an example.
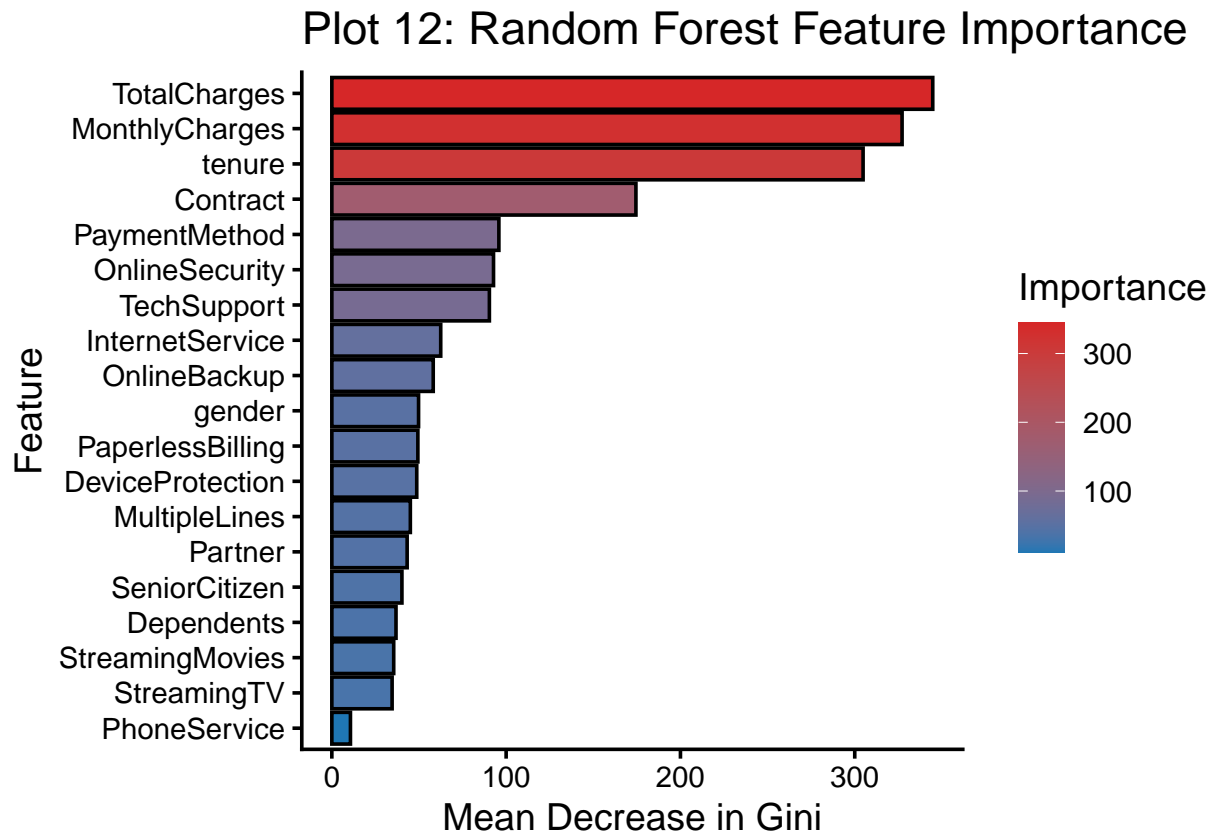
```
### Extract importance and convert into data frame
rf_imp <- importance(rf_model)

rf_imp_df <- data.frame(
  Feature = rownames(rf_imp),
  Importance = rf_imp[, 1]     # MeanDecreaseGini
)

### Rank importance
rf_imp_df <- rf_imp_df %>% arrange(desc(Importance))


### Plot output
```

```
ggplot(rf_imp_df, aes(x = reorder(Feature, Importance),
                      y = Importance,
                      fill = Importance)) +
  geom_col(color = "black") +
  coord_flip() +
  scale_fill_gradient(low="#1f77b4", high="#d62728") +
  labs(title = "Plot 12: Random Forest Feature Importance",
       x = "Feature",
       y = "Mean Decrease in Gini") +
  theme_classic(base_size = 14)
```

Plot 12: Random Forest Feature Importance

Plot 12:

- The random forest identifies TotalCharges, MonthlyCharges, and tenure as the most important predictors of churn, indicating that customer spending and length of relationship are the dominant drivers of churn behaviour.

- Contract type is also influential, while demographic and add-on service variables play a relatively smaller role.

In this case, we can make suggestions to the Telecom company about how to reduce customer churn rate:

- Focus on high-charge customers and implement differentiated pricing or promotional strategies. TotalCharges and MonthlyCharges are the most influential predictors of churn, indicating that customers with higher charges are more likely to develop churn intentions. The company may consider offering personalised discounts, bill reminders, loyalty rewards, or instalment payment options to high-spending customers in order to reduce the risk of churn caused by price pressure.

19

- Strengthen customer service support and improve the user experience. Service-quality-related variables such as TechSupport and OnlineSecurity rank highly in feature importance, demonstrating that service experience and technical support quality have a significant impact on customer retention. The company should improve response times, enhance technical support capabilities, and proactively provide services such as fault diagnosis and network optimisation to increase customer satisfaction.

- Encourage customers to adopt long-term contracts to improve customer stability. Contract type shows relatively high importance, and short-term (month-to-month) customers are generally more prone to churn. Companies should promote long-term contracts (such as one- or two-year plans) and offer additional incentives, including lower monthly fees, dedicated customer service, or device discounts, to increase customer loyalty and reduce the overall churn rate.

4.3 Limitations of the Analysis