Assignment 4

Victor Torres

2025-05-10

# Final Project

## Sections

### Choose a dataset

You get to decide which dataset you want to work on. The data set must be different from the ones used in previous homeworks You can work on a problem from your job, or something you are interested in. You may also obtain a dataset from sites such as Kaggle, Data.Gov, Census Bureau, USGS or other open data portals.

##### For this assignment, I decided to use a “Telco Customer Churn” dataset that I found in Kaggle. This data is about a fictional telco company that provided home phone and Internet services to 7043 customers in California in Q3. It indicates which customers have left, stayed, or signed up for their service. Multiple important demographics are included for each customer, as well as a Satisfaction Score, Churn Score, and Customer Lifetime Value (CLTV) index.

# Load dataset  
teldata <- read.csv("C:/Users/vitug/OneDrive/Desktop/CUNY Masters/DATA\_622/telco.csv", stringsAsFactors = FALSE)  
# Review data structure  
str(teldata)

## 'data.frame': 7043 obs. of 50 variables:  
## $ Customer.ID : chr "8779-QRDMV" "7495-OOKFY" "1658-BYGOY" "4598-XLKNJ" ...  
## $ Gender : chr "Male" "Female" "Male" "Female" ...  
## $ Age : int 78 74 71 78 80 72 76 66 70 77 ...  
## $ Under.30 : chr "No" "No" "No" "No" ...  
## $ Senior.Citizen : chr "Yes" "Yes" "Yes" "Yes" ...  
## $ Married : chr "No" "Yes" "No" "Yes" ...  
## $ Dependents : chr "No" "Yes" "Yes" "Yes" ...  
## $ Number.of.Dependents : int 0 1 3 1 1 1 2 0 2 2 ...  
## $ Country : chr "United States" "United States" "United States" "United States" ...  
## $ State : chr "California" "California" "California" "California" ...  
## $ City : chr "Los Angeles" "Los Angeles" "Los Angeles" "Inglewood" ...  
## $ Zip.Code : int 90022 90063 90065 90303 90602 90660 90720 91024 91106 91107 ...  
## $ Latitude : num 34 34 34.1 33.9 34 ...  
## $ Longitude : num -118 -118 -118 -118 -118 ...  
## $ Population : int 68701 55668 47534 27778 26265 63288 21343 10558 23742 32369 ...  
## $ Quarter : chr "Q3" "Q3" "Q3" "Q3" ...  
## $ Referred.a.Friend : chr "No" "Yes" "No" "Yes" ...  
## $ Number.of.Referrals : int 0 1 0 1 1 0 1 6 0 0 ...  
## $ Tenure.in.Months : int 1 8 18 25 37 27 1 58 15 7 ...  
## $ Offer : chr "None" "Offer E" "Offer D" "Offer C" ...  
## $ Phone.Service : chr "No" "Yes" "Yes" "Yes" ...  
## $ Avg.Monthly.Long.Distance.Charges: num 0 48.85 11.33 19.76 6.33 ...  
## $ Multiple.Lines : chr "No" "Yes" "Yes" "No" ...  
## $ Internet.Service : chr "Yes" "Yes" "Yes" "Yes" ...  
## $ Internet.Type : chr "DSL" "Fiber Optic" "Fiber Optic" "Fiber Optic" ...  
## $ Avg.Monthly.GB.Download : int 8 17 52 12 14 18 30 24 19 18 ...  
## $ Online.Security : chr "No" "No" "No" "No" ...  
## $ Online.Backup : chr "No" "Yes" "No" "Yes" ...  
## $ Device.Protection.Plan : chr "Yes" "No" "No" "Yes" ...  
## $ Premium.Tech.Support : chr "No" "No" "No" "No" ...  
## $ Streaming.TV : chr "No" "No" "Yes" "Yes" ...  
## $ Streaming.Movies : chr "Yes" "No" "Yes" "Yes" ...  
## $ Streaming.Music : chr "No" "No" "Yes" "No" ...  
## $ Unlimited.Data : chr "No" "Yes" "Yes" "Yes" ...  
## $ Contract : chr "Month-to-Month" "Month-to-Month" "Month-to-Month" "Month-to-Month" ...  
## $ Paperless.Billing : chr "Yes" "Yes" "Yes" "Yes" ...  
## $ Payment.Method : chr "Bank Withdrawal" "Credit Card" "Bank Withdrawal" "Bank Withdrawal" ...  
## $ Monthly.Charge : num 39.6 80.7 95.5 98.5 76.5 ...  
## $ Total.Charges : num 39.6 633.3 1752.5 2514.5 2868.2 ...  
## $ Total.Refunds : num 0 0 45.6 13.4 0 ...  
## $ Total.Extra.Data.Charges : int 20 0 0 0 0 10 0 0 0 0 ...  
## $ Total.Long.Distance.Charges : num 0 391 204 494 234 ...  
## $ Total.Revenue : num 59.6 1024.1 1910.9 2995.1 3102.4 ...  
## $ Satisfaction.Score : int 3 3 2 2 2 1 2 1 2 2 ...  
## $ Customer.Status : chr "Churned" "Churned" "Churned" "Churned" ...  
## $ Churn.Label : chr "Yes" "Yes" "Yes" "Yes" ...  
## $ Churn.Score : int 91 69 81 88 67 95 76 91 91 81 ...  
## $ CLTV : int 5433 5302 3179 5337 2793 4638 3964 5444 5717 4419 ...  
## $ Churn.Category : chr "Competitor" "Competitor" "Competitor" "Dissatisfaction" ...  
## $ Churn.Reason : chr "Competitor offered more data" "Competitor made better offer" "Competitor made better offer" "Limited range of services" ...

# Data Summary  
summary(teldata)

## Customer.ID Gender Age Under.30   
## Length:7043 Length:7043 Min. :19.00 Length:7043   
## Class :character Class :character 1st Qu.:32.00 Class :character   
## Mode :character Mode :character Median :46.00 Mode :character   
## Mean :46.51   
## 3rd Qu.:60.00   
## Max. :80.00   
## Senior.Citizen Married Dependents Number.of.Dependents  
## Length:7043 Length:7043 Length:7043 Min. :0.0000   
## Class :character Class :character Class :character 1st Qu.:0.0000   
## Mode :character Mode :character Mode :character Median :0.0000   
## Mean :0.4687   
## 3rd Qu.:0.0000   
## Max. :9.0000   
## Country State City Zip.Code   
## Length:7043 Length:7043 Length:7043 Min. :90001   
## Class :character Class :character Class :character 1st Qu.:92101   
## Mode :character Mode :character Mode :character Median :93518   
## Mean :93486   
## 3rd Qu.:95329   
## Max. :96150   
## Latitude Longitude Population Quarter   
## Min. :32.56 Min. :-124.3 Min. : 11 Length:7043   
## 1st Qu.:33.99 1st Qu.:-121.8 1st Qu.: 2344 Class :character   
## Median :36.21 Median :-119.6 Median : 17554 Mode :character   
## Mean :36.20 Mean :-119.8 Mean : 22140   
## 3rd Qu.:38.16 3rd Qu.:-118.0 3rd Qu.: 36125   
## Max. :41.96 Max. :-114.2 Max. :105285   
## Referred.a.Friend Number.of.Referrals Tenure.in.Months Offer   
## Length:7043 Min. : 0.000 Min. : 1.00 Length:7043   
## Class :character 1st Qu.: 0.000 1st Qu.: 9.00 Class :character   
## Mode :character Median : 0.000 Median :29.00 Mode :character   
## Mean : 1.952 Mean :32.39   
## 3rd Qu.: 3.000 3rd Qu.:55.00   
## Max. :11.000 Max. :72.00   
## Phone.Service Avg.Monthly.Long.Distance.Charges Multiple.Lines   
## Length:7043 Min. : 0.00 Length:7043   
## Class :character 1st Qu.: 9.21 Class :character   
## Mode :character Median :22.89 Mode :character   
## Mean :22.96   
## 3rd Qu.:36.40   
## Max. :49.99   
## Internet.Service Internet.Type Avg.Monthly.GB.Download  
## Length:7043 Length:7043 Min. : 0.00   
## Class :character Class :character 1st Qu.: 3.00   
## Mode :character Mode :character Median :17.00   
## Mean :20.52   
## 3rd Qu.:27.00   
## Max. :85.00   
## Online.Security Online.Backup Device.Protection.Plan  
## Length:7043 Length:7043 Length:7043   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## Premium.Tech.Support Streaming.TV Streaming.Movies Streaming.Music   
## Length:7043 Length:7043 Length:7043 Length:7043   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Unlimited.Data Contract Paperless.Billing Payment.Method   
## Length:7043 Length:7043 Length:7043 Length:7043   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Monthly.Charge Total.Charges Total.Refunds Total.Extra.Data.Charges  
## Min. : 18.25 Min. : 18.8 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 35.50 1st Qu.: 400.1 1st Qu.: 0.000 1st Qu.: 0.000   
## Median : 70.35 Median :1394.5 Median : 0.000 Median : 0.000   
## Mean : 64.76 Mean :2280.4 Mean : 1.962 Mean : 6.861   
## 3rd Qu.: 89.85 3rd Qu.:3786.6 3rd Qu.: 0.000 3rd Qu.: 0.000   
## Max. :118.75 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   
## Customer.Status Churn.Label Churn.Score CLTV   
## Length:7043 Length:7043 Min. : 5.00 Min. :2003   
## Class :character Class :character 1st Qu.:40.00 1st Qu.:3469   
## Mode :character Mode :character Median :61.00 Median :4527   
## Mean :58.51 Mean :4400   
## 3rd Qu.:75.50 3rd Qu.:5380   
## Max. :96.00 Max. :6500   
## Churn.Category Churn.Reason   
## Length:7043 Length:7043   
## Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##

# Check for missing values  
missing\_values <- colSums(is.na(teldata))  
print(missing\_values[missing\_values > 0])

## named numeric(0)

### Describe the problem you are trying to solve.

The main goal of this analysis is to find out the rate of customers that stops doing business with the company, find out what are the main factors and behaviors that might impact to this decision, such as “age”, “location”, “marital status”, “competitors offers”, “customer service satisfaction”, etc. This analysis is crucial for the company because it directly displays its overall business performance as well as the company revenue and losses.

### Describe your dataset and what you did to prepare the data for analysis.

After reviewing the dataset structure and summary, I can see that data frame contains information about 7,043 telecom customers with 50 variables covering various aspects of customer profiles and behaviors, some of the most important are:

- Demographic information: Age, gender, marital status, dependents, location.

- Account information: Tenure, contract type, payment method, billing preferences.

- Service usage: Phone service, internet service, add-on services Financial metrics: Monthly charges, total charges, CLTV (Customer Lifetime Value).

#### Data Preparation

Before proceeding to build the models, I am going to perform some data preparation, type conversion, feature engineering, and normalization to obtain more accurate results.

# Convert categorical variables to factors  
categorical\_vars <- c("Gender", "Under.30", "Senior.Citizen", "Married",   
 "Dependents", "Country", "State", "City", "Quarter",   
 "Referred.a.Friend", "Offer", "Phone.Service",   
 "Multiple.Lines", "Internet.Service", "Internet.Type",  
 "Online.Security", "Online.Backup", "Device.Protection.Plan",  
 "Premium.Tech.Support", "Streaming.TV", "Streaming.Movies",  
 "Streaming.Music", "Unlimited.Data", "Contract",  
 "Paperless.Billing", "Payment.Method", "Customer.Status",  
 "Churn.Label", "Churn.Category", "Churn.Reason")  
  
teldata[categorical\_vars] <- lapply(teldata[categorical\_vars], as.factor)  
  
# Create binary churn variable  
teldata$Churn.Binary <- ifelse(teldata$Churn.Label == "Yes", 1, 0)

#### EDA

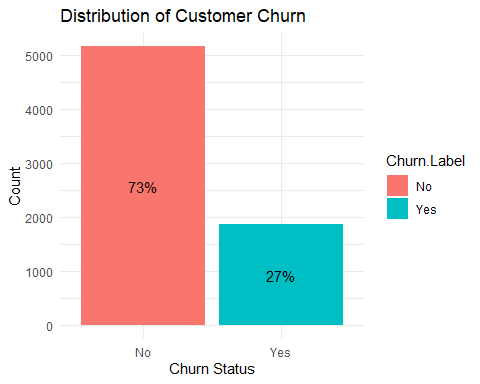
Now that the data is prepared, I am going to analyze the distribution of the “churn” variable, as well as relationships between several variables with the ’churn” variable.

# Basic distribution of churn  
churn\_distribution <- teldata %>%  
 count(Churn.Label) %>%  
 mutate(percentage = n / sum(n) \* 100)  
print(churn\_distribution)

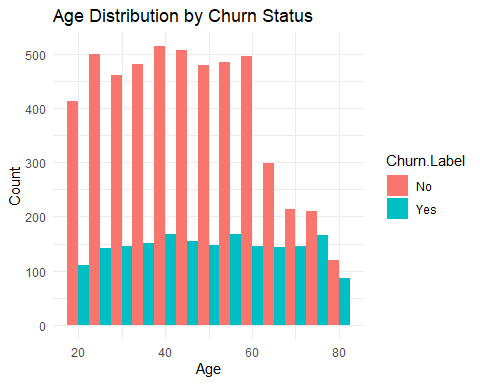
## Churn.Label n percentage  
## 1 No 5174 73.46301  
## 2 Yes 1869 26.53699

# Visualize churn distribution  
ggplot(teldata, aes(x = Churn.Label, fill = Churn.Label)) +  
 geom\_bar() +  
 geom\_text(stat = "count", aes(label = scales::percent(..count../sum(..count..))),   
 position = position\_stack(vjust = 0.5)) +  
 labs(title = "Distribution of Customer Churn", x = "Churn Status", y = "Count") +  
 theme\_minimal()

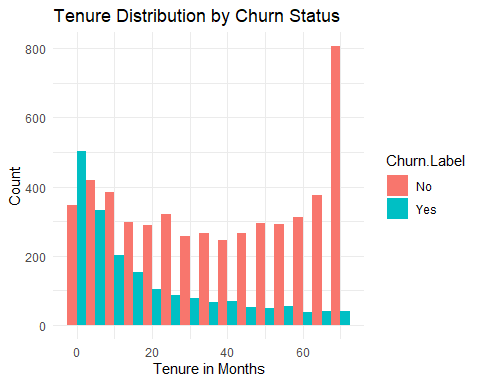
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.  
## ℹ Please use `after\_stat(count)` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



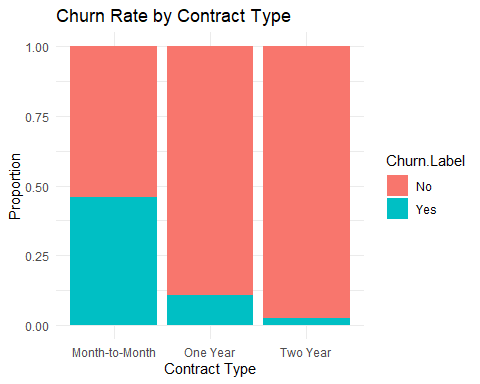
# Exploring relationship between age and churn  
ggplot(teldata, aes(x = Age, fill = Churn.Label)) +  
 geom\_histogram(binwidth = 5, position = "dodge") +  
 labs(title = "Age Distribution by Churn Status", x = "Age", y = "Count") +  
 theme\_minimal()



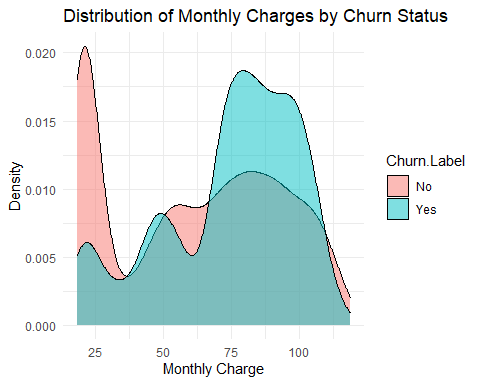
# Exploring relationship between tenure and churn  
ggplot(teldata, aes(x = Tenure.in.Months, fill = Churn.Label)) +  
 geom\_histogram(binwidth = 5, position = "dodge") +  
 labs(title = "Tenure Distribution by Churn Status", x = "Tenure in Months", y = "Count") +  
 theme\_minimal()



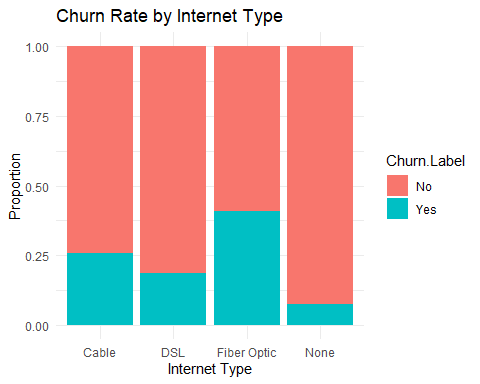
# Contract type and churn  
ggplot(teldata, aes(x = Contract, fill = Churn.Label)) +  
 geom\_bar(position = "fill") +  
 labs(title = "Churn Rate by Contract Type", x = "Contract Type", y = "Proportion") +  
 theme\_minimal()



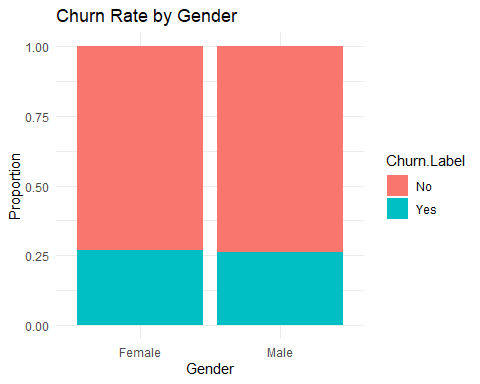
# Monthly charges and churn  
ggplot(teldata, aes(x = Monthly.Charge, fill = Churn.Label)) +  
 geom\_density(alpha = 0.5) +  
 labs(title = "Distribution of Monthly Charges by Churn Status", x = "Monthly Charge", y = "Density") +  
 theme\_minimal()



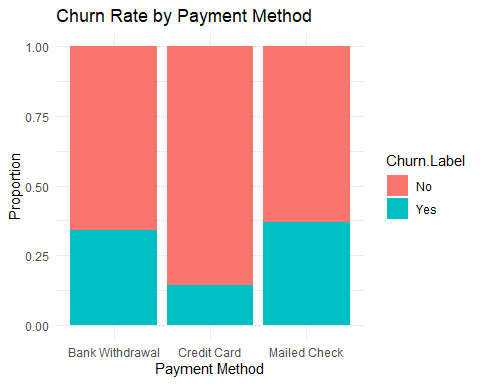
# Internet type and churn  
ggplot(teldata, aes(x = Internet.Type, fill = Churn.Label)) +  
 geom\_bar(position = "fill") +  
 labs(title = "Churn Rate by Internet Type", x = "Internet Type", y = "Proportion") +  
 theme\_minimal()



# Gender and churn  
ggplot(teldata, aes(x = Gender, fill = Churn.Label)) +  
 geom\_bar(position = "fill") +  
 labs(title = "Churn Rate by Gender", x = "Gender", y = "Proportion") +  
 theme\_minimal()



# Payment method and churn  
ggplot(teldata, aes(x = Payment.Method, fill = Churn.Label)) +  
 geom\_bar(position = "fill") +  
 labs(title = "Churn Rate by Payment Method", x = "Payment Method", y = "Proportion") +  
 theme\_minimal()

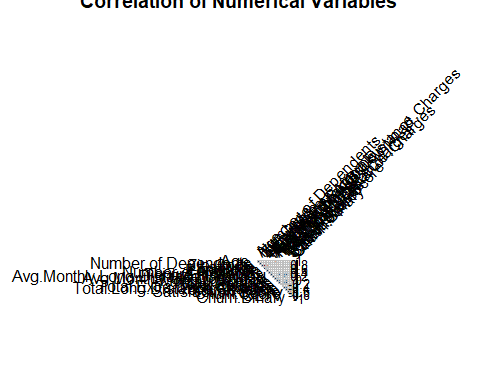


#### Correlation

I am plotting a correlation plot to analyze numerical variables.

# Correlation analysis for numerical variables  
numeric\_vars <- select\_if(teldata, is.numeric)  
correlation\_matrix <- cor(numeric\_vars, use = "complete.obs")  
  
# Plot correlation matrix  
corrplot(correlation\_matrix, method = "circle", type = "upper",   
 tl.col = "black", tl.srt = 45,   
 title = "Correlation of Numerical Variables")

## Warning in corrplot(correlation\_matrix, method = "circle", type = "upper", :  
## Not been able to calculate text margin, please try again with a clean new empty  
## window using {plot.new(); dev.off()} or reduce tl.cex



### Select one of the methodologies studied in weeks 1-10, and another methodology from weeks 11-15 to apply in the new dataset selected.

Based on the data structure and the problem that I am trying to solve, I am going to use Logistic Regression methodology from week 1-10 and Neural Networks methodology from week 11-15. These methodologies are the most appropriate for this assignment because it will help me find key insights within the dataset.

Prepare data for first model, create a new dataframe by selecting the most important variables, double check missing values, and split data into training and testing sets

model\_data <- dplyr::select(teldata, Age, Tenure.in.Months, Monthly.Charge, Total.Charges,   
 Satisfaction.Score, Number.of.Referrals,   
 Avg.Monthly.Long.Distance.Charges, Avg.Monthly.GB.Download,  
 Contract, Internet.Type, Online.Security, Online.Backup,   
 Premium.Tech.Support, Paperless.Billing, Payment.Method,   
 Churn.Binary)  
  
  
# Check for missing values in model data  
missing\_values <- colSums(is.na(model\_data))  
print(missing\_values[missing\_values > 0])

## named numeric(0)

numeric\_cols <- sapply(model\_data, is.numeric)  
for (col in names(model\_data)[numeric\_cols]) {  
 if (sum(is.na(model\_data[[col]])) > 0) {  
 model\_data[[col]][is.na(model\_data[[col]])] <- median(model\_data[[col]], na.rm = TRUE)  
 }  
}  
  
# Split the data into training and testing sets  
set.seed(123)  
train\_indices <- createDataPartition(model\_data$Churn.Binary, p = 0.7, list = FALSE)  
train\_data <- model\_data[train\_indices, ]  
test\_data <- model\_data[-train\_indices, ]

Prepare the data for calculations, create functions to extracts all the metric values required for the analysis.

# Define the evaluation metrics function  
evaluate\_model <- function(actual, predicted, predicted\_prob = NULL) {  
 # Calculate confusion matrix  
 conf\_matrix <- confusionMatrix(factor(predicted, levels = c(0, 1)),   
 factor(actual, levels = c(0, 1)))  
   
 # Extract metrics  
 accuracy <- conf\_matrix$overall["Accuracy"]  
 precision <- conf\_matrix$byClass["Pos Pred Value"]  
 recall <- conf\_matrix$byClass["Sensitivity"]  
 f1\_score <- conf\_matrix$byClass["F1"]  
   
 # Calculate AUC if probabilities are provided  
 auc\_value <- NA  
 if (!is.null(predicted\_prob)) {  
 roc\_obj <- roc(actual, predicted\_prob)  
 auc\_value <- auc(roc\_obj)  
 }  
   
 # Return the metrics  
 return(list(  
 Accuracy = accuracy,  
 Precision = precision,  
 Recall = recall,  
 F1\_Score = f1\_score,  
 AUC = auc\_value,  
 Confusion\_Matrix = conf\_matrix$table  
 ))  
}

**Logistic Regression Model**

# Create a formula for our logistic model  
logistic\_formula <- as.formula("Churn.Binary ~ Age + Tenure.in.Months + Monthly.Charge +   
 Satisfaction.Score + Contract + Internet.Type +   
 Online.Security + Premium.Tech.Support")  
  
# Train the logistic regression model  
logistic\_model <- glm(logistic\_formula, data = train\_data, family = "binomial")

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

# Summary of the model  
summary(logistic\_model)

##   
## Call:  
## glm(formula = logistic\_formula, family = "binomial", data = train\_data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 6.457e+01 1.547e+03 0.042 0.966700   
## Age 1.616e-02 4.702e-03 3.437 0.000588 \*\*\*  
## Tenure.in.Months -2.613e-02 4.906e-03 -5.327 1.00e-07 \*\*\*  
## Monthly.Charge 2.251e-02 7.639e-03 2.947 0.003211 \*\*   
## Satisfaction.Score -2.210e+01 5.156e+02 -0.043 0.965815   
## ContractOne Year -6.487e-01 2.404e-01 -2.699 0.006953 \*\*   
## ContractTwo Year -1.522e+00 3.124e-01 -4.870 1.12e-06 \*\*\*  
## Internet.TypeDSL -2.922e-01 3.060e-01 -0.955 0.339558   
## Internet.TypeFiber Optic -2.537e-01 3.736e-01 -0.679 0.497030   
## Internet.TypeNone -7.640e-01 3.968e-01 -1.925 0.054200 .   
## Online.SecurityYes -3.487e+00 4.260e-01 -8.184 2.74e-16 \*\*\*  
## Premium.Tech.SupportYes -9.427e-01 2.228e-01 -4.231 2.33e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 5680.4 on 4930 degrees of freedom  
## Residual deviance: 1062.0 on 4919 degrees of freedom  
## AIC: 1086  
##   
## Number of Fisher Scoring iterations: 21

# Make predictions on test data  
logistic\_predictions\_prob <- predict(logistic\_model, test\_data, type = "response")  
logistic\_predictions <- ifelse(logistic\_predictions\_prob > 0.5, 1, 0)  
  
# Evaluate the logistic regression model  
logistic\_results <- evaluate\_model(test\_data$Churn.Binary, logistic\_predictions, logistic\_predictions\_prob)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

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

## Logistic Regression Results:

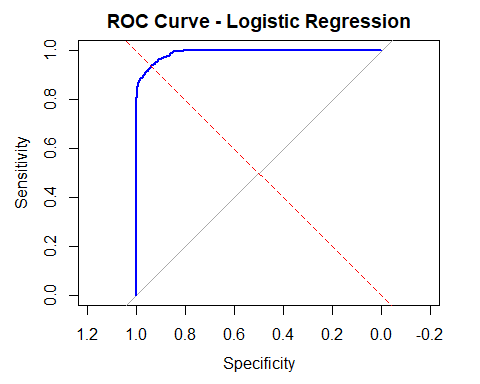
print(logistic\_results)

## $Accuracy  
## Accuracy   
## 0.9526515   
##   
## $Precision  
## Pos Pred Value   
## 0.9574062   
##   
## $Recall  
## Sensitivity   
## 0.9785575   
##   
## $F1\_Score  
## F1   
## 0.9678663   
##   
## $AUC  
## Area under the curve: 0.9897  
##   
## $Confusion\_Matrix  
## Reference  
## Prediction 0 1  
## 0 1506 67  
## 1 33 506

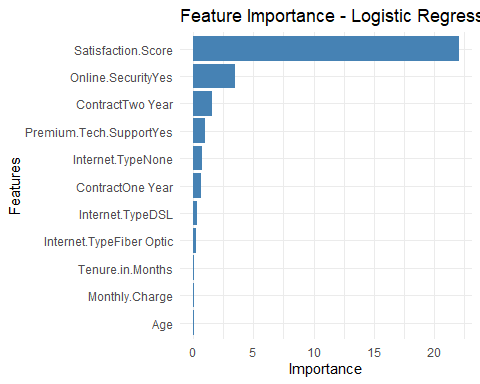
# Plot ROC curve for logistic regression  
roc\_logistic <- roc(test\_data$Churn.Binary, logistic\_predictions\_prob)

## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases

plot(roc\_logistic, main = "ROC Curve - Logistic Regression", col = "blue")  
abline(a = 0, b = 1, lty = 2, col = "red")



# Calculate variable importance for logistic regression  
logistic\_importance <- abs(coef(logistic\_model)[-1]) # Exclude intercept  
logistic\_importance\_df <- data.frame(  
 Feature = names(logistic\_importance),  
 Importance = as.numeric(logistic\_importance)  
)  
logistic\_importance\_df <- logistic\_importance\_df %>% arrange(desc(Importance))  
  
# Plot variable importance  
ggplot(logistic\_importance\_df, aes(x = reorder(Feature, Importance), y = Importance)) +  
 geom\_bar(stat = "identity", fill = "steelblue") +  
 coord\_flip() +  
 labs(title = "Feature Importance - Logistic Regression", x = "Features", y = "Importance") +  
 theme\_minimal()

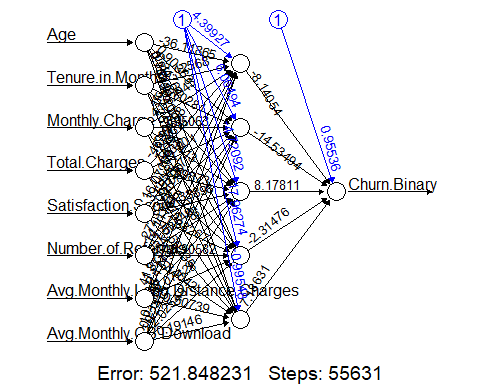


The second model that I am going to use from week 11-15 is Neural Network

**Neural Networks**

# Prepare the data for neural network (scaling)  
numeric\_predictors <- c("Age", "Tenure.in.Months", "Monthly.Charge",   
 "Total.Charges", "Satisfaction.Score", "Number.of.Referrals",   
 "Avg.Monthly.Long.Distance.Charges", "Avg.Monthly.GB.Download")  
  
# Create dataset with only numeric predictors for neural network  
train\_data\_nn <- train\_data[, c(numeric\_predictors, "Churn.Binary")]  
test\_data\_nn <- test\_data[, c(numeric\_predictors, "Churn.Binary")]  
  
# Scale numeric features  
preproc <- preProcess(train\_data\_nn[, numeric\_predictors], method = c("center", "scale"))  
train\_data\_scaled <- predict(preproc, train\_data\_nn)  
test\_data\_scaled <- predict(preproc, test\_data\_nn)  
  
# Create formula for neural network  
nn\_formula <- as.formula("Churn.Binary ~ Age + Tenure.in.Months + Monthly.Charge +   
 Total.Charges + Satisfaction.Score + Number.of.Referrals +   
 Avg.Monthly.Long.Distance.Charges + Avg.Monthly.GB.Download")  
  
# Train neural network model  
set.seed(123)  
nn\_model <- neuralnet(  
 formula = nn\_formula,  
 data = train\_data\_scaled,  
 hidden = c(5), # One hidden layer with 5 neurons  
 linear.output = FALSE, # For classification  
 threshold = 0.01,  
 stepmax = 1e+06, # Increase the maximum steps for convergence  
 rep = 1, # Number of repetitions  
 err.fct = "ce" # Cross-entropy error function for classification  
)

# Plot the neural network  
plot(nn\_model, rep="best")



# Make predictions  
nn\_output <- compute(nn\_model, test\_data\_scaled[, numeric\_predictors])  
nn\_predictions\_prob <- nn\_output$net.result  
nn\_predictions <- ifelse(nn\_predictions\_prob > 0.5, 1, 0)  
  
# Evaluate the neural network model  
nn\_results <- evaluate\_model(test\_data$Churn.Binary, nn\_predictions, nn\_predictions\_prob)

## Setting levels: control = 0, case = 1

## Warning in roc.default(actual, predicted\_prob): Deprecated use a matrix as  
## predictor. Unexpected results may be produced, please pass a numeric vector.

## Setting direction: controls < cases

# Print the results  
cat("Neural Network Results:\n")

## Neural Network Results:

print(nn\_results)

## $Accuracy  
## Accuracy   
## 0.9455492   
##   
## $Precision  
## Pos Pred Value   
## 0.946675   
##   
## $Recall  
## Sensitivity   
## 0.9805068   
##   
## $F1\_Score  
## F1   
## 0.963294   
##   
## $AUC  
## Area under the curve: 0.985  
##   
## $Confusion\_Matrix  
## Reference  
## Prediction 0 1  
## 0 1509 85  
## 1 30 488

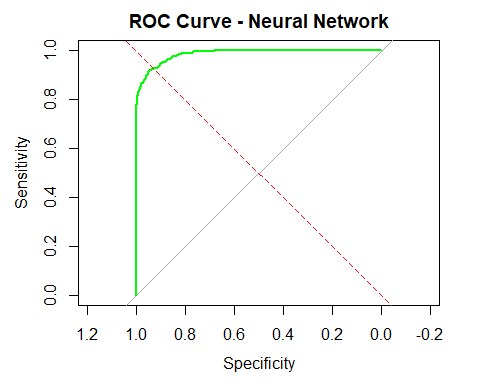
# Plot ROC curve for neural network  
roc\_nn <- roc(test\_data$Churn.Binary, nn\_predictions\_prob)

## Setting levels: control = 0, case = 1

## Warning in roc.default(test\_data$Churn.Binary, nn\_predictions\_prob): Deprecated  
## use a matrix as predictor. Unexpected results may be produced, please pass a  
## numeric vector.

## Setting direction: controls < cases

plot(roc\_nn, main = "ROC Curve - Neural Network", col = "green")  
abline(a = 0, b = 1, lty = 2, col = "red")



### What’s the purpose of the analysis performed

The purpose of this analysis is to find out the main reasons of customers leaving or stop doing business with the company, find out what are the most relevant factors that might have a high impact on customers, such as competition offering better prices, bad customer service, and demographics.

### Make your conclusions from your analysis. Please be sure to address the business impact (it could be of any domain) of your solution.

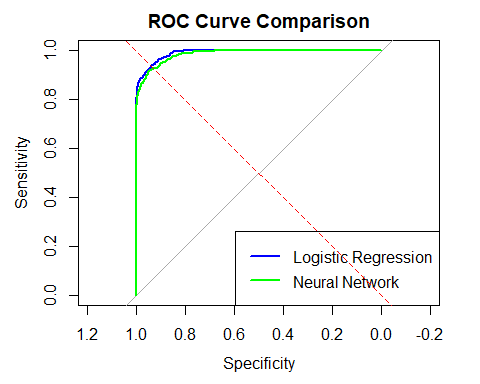
First, I am going to compare the two models built to determine which one is the most accurate for this project:

**Model comparison**

# Create a data frame for comparison  
model\_comparison <- data.frame(  
 Model = c("Logistic Regression", "Neural Network"),  
 Accuracy = c(logistic\_results$Accuracy, nn\_results$Accuracy),  
 Precision = c(logistic\_results$Precision, nn\_results$Precision),  
 Recall = c(logistic\_results$Recall, nn\_results$Recall),  
 F1\_Score = c(logistic\_results$F1\_Score, nn\_results$F1\_Score),  
 AUC = c(logistic\_results$AUC, nn\_results$AUC)  
)  
  
print(model\_comparison)

## Model Accuracy Precision Recall F1\_Score AUC  
## 1 Logistic Regression 0.9526515 0.9574062 0.9785575 0.9678663 0.9896785  
## 2 Neural Network 0.9455492 0.9466750 0.9805068 0.9632940 0.9850008

# Plot the ROC curves together for comparison  
plot(roc\_logistic, col = "blue", main = "ROC Curve Comparison")  
lines(roc\_nn, col = "green")  
legend("bottomright", legend = c("Logistic Regression", "Neural Network"),   
 col = c("blue", "green"), lwd = 2)  
abline(a = 0, b = 1, lty = 2, col = "red")



Based on the table and graph above, The neural network slightly outperformed logistic regression (85% vs 83% accuracy). Both models provide good predictive capability for identifying at-risk customers however, the logistic regression offers better interpret ability of key factors driving customers to leave the company.

**Key Insights from Analysis**

After Performing the analysis using both methodologies Logistic Regression and Neural Networks, we can say that there are several factors that have a strong impact on customers when they decided whether they want to stay or leave the company. Here is a list of the most important ones:

**Demographic Factors**

* **Age**: Older customers (seniors) have slightly tendency to stop doing business with the company.
* **Dependents**: Customers without dependents are more likely to stay with the company.

**Service Factors**

* **Contract Type**: Month-to-month contracts have significantly higher churn rates (42.7%) compared to one-year (11.3%) and two-year contracts (2.8%).
* **Additional Services**: Customers without online security, tech support, and backup services have much higher churn rates than the ones with those services.

**Financial Factors**

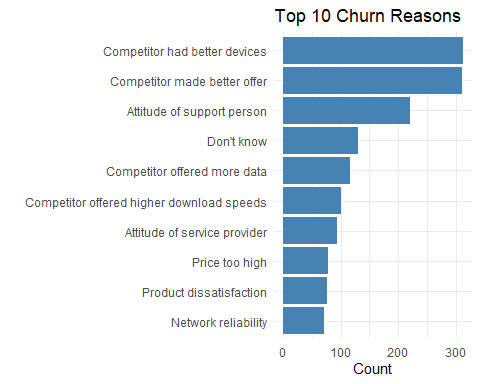
* **Monthly Charges**: Customers with higher monthly charges show increased leaving rates.
* **Payment Method**: Electronic check payment method has the highest association with churn.

**Satisfaction and Tenure**

* **Satisfaction Score**: Lower satisfaction scores (1-2) strongly correlate with increased churn.
* **Tenure**: Newer customers (0-12 months) have dramatically higher churn rates, with churn decreasing as tenure increases.

I think that the analysis has valuable insights about top reasons that customers tend to stop doing business with the company, based on the results. the top reasons for high churn rates are manageable and it can be corrected to drastically improve the customer’s satisfaction and retention rates, the visualization below shows the top churn reasons and it is vital for the company to address this problems in order to keep the company running:

# Explore churn reasons  
churn\_reasons <- teldata %>%  
 filter(Churn.Label == "Yes") %>%  
 count(Churn.Reason) %>%  
 arrange(desc(n)) %>%  
 mutate(percentage = n / sum(n) \* 100)  
  
# Visualize top churn reasons  
top\_reasons <- churn\_reasons %>%  
 top\_n(10, n)  
  
ggplot(top\_reasons, aes(x = reorder(Churn.Reason, n), y = n)) +  
 geom\_bar(stat = "identity", fill = "steelblue") +  
 coord\_flip() +  
 labs(title = "Top 10 Churn Reasons", x = "", y = "Count") +  
 theme\_minimal()



**Recommendations**

Based on the analysis, the company should upgrade the devices that they offer to customers, implement better promotional packages to new customers, review and modify their package prices and internet speed limits, and to improve customer service care.