Machine Learning

Assignment 1

PURPOSE:

The purpose of this assignment is to set up R (R Studio) and use its basic commands.

ABOUT THE DATASET:

Churn refers to the phenomenon where customers stop using a service or product, often switching to a competitor. In the context of mobile phone usage, churn is a critical issue for telecom companies as it directly impacts revenue and market share. Understanding the factors that contribute to customer churn allows companies to devise strategies to retain their users.

Here are the one-line definitions with their impact on churn:

1. **Call Failures**: Number of failed call attempts; frequent failures can increase churn due to dissatisfaction.
2. **Subscription Length**: Total time with the service; longer subscriptions usually indicate loyalty but may not prevent churn.
3. **Data Usage**: Amount of data used; extreme usage patterns can signal potential churn.
4. **Voice Minutes**: Total voice call time; sudden drops may indicate dissatisfaction and risk of churn.
5. **Customer Support Calls**: Number of support interactions; frequent calls often signal dissatisfaction and higher churn risk.
6. **Contract Type**: Type of service contract; flexible contracts like month-to-month increase churn likelihood.
7. **Monthly Charges**: Customer's monthly bill; high charges without perceived value can lead to churn.
8. **Roaming Usage**: Usage outside the home network; high roaming charges may push customers towards churn.
9. **CHURN**: Binary indicator of whether a customer has left the service (1) or not (0); the primary target variable for churn analysis.

This assignment will concentrate on using R. You must first import the provided dataset and install all necessary packages such as dplyr, tidyverse, ggplot2, psych, tinytex, readr.

1. [20 Points] Conduct basic descriptive statistics, including head, tail, dimensions, summary, structure, mean, mode, and median, and check for any missing values.
   Hint: Necessary commands include str, dim, summary, mean, mode, median, is.na, etc.}
2. [20 Points] Write an interpretation of your analysis in part 1 (descriptive statistics) in R Markdown report.
3. [15 Points] Select and compare two or more variables from the dataset, such as **Data Usage** and **CHURN**. Create a table for the selected variables using the df(dataframe) and select commands.
   Hint: In your comparison, explain how **Data Usage** relates to customer churn. For example, analyze whether higher data usage is associated with a higher or lower likelihood of churn. Provide summary statistics or visualizations to support your comparison.
4. [20 Points] Transform at least one variable (any transformation such as log(x), sqrt(x), scale(x) is acceptable) and generate a plot using 'ggplot' for the transformed variable.
5. Please submit your answer in RMD **and** pdf/html.

File Attached: **Churn.csv**