D209 Performance Assessment

Data Mining I

Task 2 Predictive Analysis

Instructor: Dr. Elleh

Student Name: Nicole Haibach

Student ID: 001260374

Email: nhaibac@wgu.edu

# Research Question

## Research Question

The research question for this assignment will look at the churn of a telecommunication company. How do the different factors affect who churn compared to those who remain with the telecommunications company, and can decision tree analysis be utilized to predict customer churn based on these attributes within a dataset of 10,000 customers?

## Goal of Analysis

The objective of utilizing a decision tree analysis is to uncover the key factors that determine the success or failure of our intervention strategy to prevent churn of customers. Through the systematic examination of our dataset using decision trees, we aim to identify the critical predictors driving the effectiveness of our intervention. By understanding the decision rules and patterns within the data, we seek to gain insights that will inform strategic decision-making and refinement of our intervention approach. Ultimately, the goal is to leverage these insights to optimize outcomes and enhance the impact of our intervention efforts.

# Method Justification

## Benefits of using KNN

The decision tree analysis method systematically evaluates the selected dataset by recursively splitting it into smaller subsets based on predictor variables, aiming to uncover patterns and relationships. It identifies key predictors that contribute to the target variable and generates decision rules describing how different variables influence outcomes. The expected outcomes include improved understanding of dataset dynamics, enhanced predictive accuracy, and actionable insights for decision-making and strategy development. Ultimately, the analysis aims to inform targeted interventions and drive improvements in the targeted outcome.

## Summarize Assumptions

One assumption of the decision tree analysis method is that it assumes the independence of predictor variables. In other words, the algorithm assumes that each predictor variable's value is not influenced by or correlated with the values of other predictor variables. This assumption ensures that the decision tree accurately captures the relationships between predictors and the target variable without being confounded by multicollinearity or dependencies among predictors. However, in real-world datasets, predictor variables may be correlated, and violating this assumption can affect the accuracy and interpretability of the decision tree model.

## Justify Technique

1. **readxl**: It supports the analysis by enabling the loading of the dataset containing customer data, which is crucial for conducting the churn analysis.
2. **visdat**: This library provides visualization tools for exploring missing values in the dataset. It supports the analysis by allowing the identification and handling of missing data, ensuring the dataset's completeness and quality for further analysis.
3. **dplyr**: This library provides a set of functions for data manipulation and transformation. It supports the analysis by facilitating tasks such as filtering, summarizing, and transforming the dataset to prepare it for modeling, including feature engineering and data preprocessing.
4. **ggplot2**: **ggplot2** is a powerful data visualization package in R, allowing the creation of customizable and publication-quality plots. It supports the analysis by facilitating the visualization of relationships between variables, model diagnostics, and performance metrics, aiding in data exploration and model evaluation.
5. **caret**: This package in R provides a unified interface for training and tuning predictive models. It offers a wide range of functions for preprocessing data, model training, parameter tuning, and model evaluation.
6. **rpart**: This package provides functionality for building and visualizing decision trees using the Recursive Partitioning and Regression Trees (rpart) algorithm. It supports the analysis by allowing the construction of decision tree models to uncover patterns and relationships in the data, thereby identifying key predictors and generating decision rules.
7. **rpart.plot**: This0 is a package specifically designed for visualizing decision trees created using the rpart package in R. Decision trees are a popular machine learning algorithm for classification and regression tasks due to their simplicity and interpretability.

These packages support the analysis by providing tools for data preprocessing, model building, evaluation, and visualization, thereby enabling a comprehensive approach to customer churn prediction using machine learning techniques in R.

# Data Preparation

## Data Processing Goal

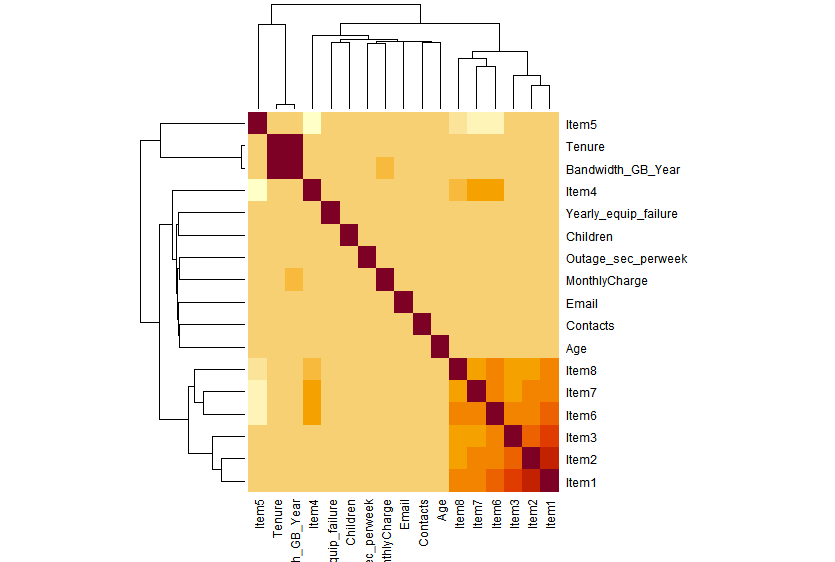
One data preprocessing goal relevant to the decision tree prediction method is handling missing values in the dataset. Missing data can negatively impact the performance of decision tree models by introducing bias and reducing predictive accuracy. Therefore, the goal is to implement strategies to effectively deal with missing values, such as imputation or deletion of records with missing values. The choice of imputation method should be carefully considered to minimize bias and preserve data integrity, ensuring that the decision tree model is trained on a complete and representative dataset for more reliable predictions.

## Variables

| **Variable Name** | **Data Type** | **Description** | **Example** |
| --- | --- | --- | --- |
| Area | Categorical | Type of area based on census | Suburban |
| Children | Continuous | The number of children the customer has | 2 |
| Age | Continuous | Age of Customer | 49 |
| Marital | Categorical | Customers marital status | Separated |
| Gender | Categorical | Customer Gender identified by the customer | Prefer not to answer |
| Churn | Categorical | If customers discontinue services | No |
| Outage\_sec\_perweek | Continuous | Average seconds of outages per week in the area | 6.637258801 |
| Email | Continuous | Emails sent to customers last year | 20 |
| Contacts | Continuous | Number of Technical support contacts | 2 |
| Yearly\_equip\_failure | Continuous | Number of equipment failures within a year | 3 |
| Techie | Categorical | If customers feel technically inclined | NA |
| Contract | Categorical | Length of contract | Month-to-month |
| Port\_modem | Categorical | If the customer has a portable modem | Yes |
| Tablet | Categorical | If the customer has a tablet | No |
| InternetService | Categorical | Internet provider | DSL |
| Phone | Categorical | If the customer has phone services | Yes |
| Multiple | Categorical | If the customer has multiple lines | No |
| OnlineSecurity | Categorical | If the customer has security for their computer | Yes |
| Online backup | Categorical | If the customer keeps a backup | Yes |
| DeviceProtection | Categorical | If the customer has the protection add-on | No |
| TechSupport | Categorical | If the customer has a technical support add-on | No |
| StreamingTV | Categorical | If the customer has a streaming television | No |
| StreamingMovies | Categorical | If the customer has streaming movies | No |
| PaperlessBilling | Categorical | If the customer opted for paperless billing | Yes |
| payment method | Categorical | The method of payment chosen by the customer | Bank Transfer(automatic) |
| Tenure | Continuous | length of time the customer has stayed with the provider in months | 8.220686373 |
| MonthlyCharge | Continuous | The average amount charged to customers per month | 118.3668439 |
| Bandwith\_GB\_Year | Continuous | Average GBs of data used by customers per year | 1312.874964 |
| Item1 | Categorical | Importance scale rating for timely responses | 5 |
| Item2 | Categorical | Importance scale rating for timely fixes | 4 |
| Item3 | Categorical | Importance scale rating for timely replacements | 4 |
| Item4 | Categorical | Importance scale rating for reliability | 3 |
| Item5 | Categorical | Importance scale rating for options | 4 |
| Item6 | Categorical | Importance scale rating for Respectful responses | 3 |
| Item7 | Categorical | Importance scale rating for courteous exchange | 4 |
| Item8 | Categorical | Importance scale rating for evidence of active listening | 4 |

## Describe cleaning steps and goals

The goal of cleaning the data is to ensure the usability of the data set. The data cleaning process begins by checking for repeated data entries and identifying missing values across the dataset, which are visualized for clarity. Unnecessary columns such as identifiers, geographical details, and irrelevant demographic information are removed to streamline the dataset. Subsequently, column names are appropriately renamed for clarity and consistency. After inspecting the data's structure and summarizing its key statistics, only numeric columns are retained for outlier detection and subsequent analysis. Outliers are visually assessed through boxplots, and correlations between numeric variables are examined using a heatmap. Finally, categorical variables are encoded into numerical format using one-hot encoding, ensuring compatibility with decision tree algorithms for subsequent analysis.



###########Data Cleaning###################

#Checking for repeated data

CC <- distinct(CC)

#Checking for missing values throughout dataset

vis\_miss(CC)

# Remove unnescessary columns

SC <- CC[, !(names(CC) %in% c("CaseOrder", "Customer\_id", "Interaction", "UID", "City", "State", "County", "Zip", "Lat", "Lng", "Population", "TimeZone", "Income", "Job"))]

#renaming Columns for a proper naming

names(SC)[names(SC) == "Item1"] <- "TimelyResponse"

names(SC)[names(SC) == "Item2"] <- "TimelyFixes"

names(SC)[names(SC) == "Item3"] <- "TimelyReplacements"

names(SC)[names(SC) == "Item4"] <- "Reliability"

names(SC)[names(SC) == "Item5"] <- "Options"

names(SC)[names(SC) == "Item6"] <- "RespectfulResponse"

names(SC)[names(SC) == "Item7"] <- "CourteousExchange"

names(SC)[names(SC) == "Item8"] <- "ActiveListening"

#Inspecting Data

str(SC)

summary(SC)

# Select only numeric columns

numeric\_SC <- SC[sapply(SC, is.numeric)]

# Check for outliers in each numeric column using boxplots

boxplot(numeric\_SC)

# Create a correlation matrix of numeric variables

correlation\_matrix <- cor(numeric\_SC)

# Create a heatmap of the correlation matrix

heatmap(correlation\_matrix, symm = TRUE, margins = c(5, 10))

# Perform one-hot encoding

SC <- predict(dummyVars(~ ., data = SC, fullRank = TRUE), newdata = SC)

## Provide a Copy of Cleaned Data Set

# Specify the file path and name for the CSV file

csv\_file\_path <- "C:/Users/nshai/OneDrive/Pictures/Documents/School/D209B/CleanedData.csv"

# Export the dataset to a CSV file

write.csv(SC, file = csv\_file\_path, row.names = FALSE)

# Print a message indicating the successful export

cat("Dataset exported to:", csv\_file\_path, "\n")

Cleaned data provided in CSV file named “CleanedData”

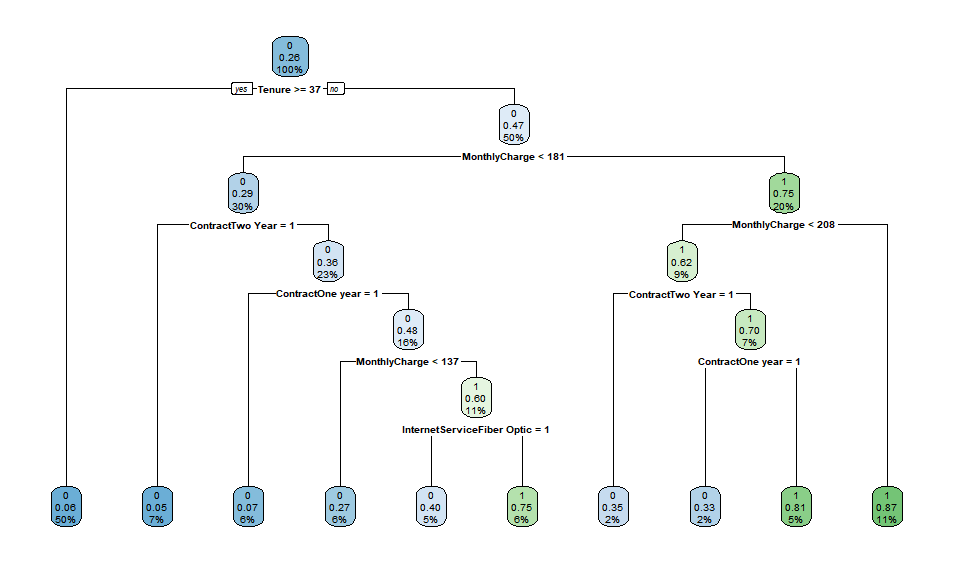
# Analysis

## Training and Test Data Sets

Data was split into training and test data sets with files provided.

## Describe Analysis Techniques

The chosen analysis technique was the decision tree algorithm, specifically CART. This method efficiently handles both categorical and numerical variables, capturing nonlinear relationships and interactions. Through partitioning, it segments the data based on predictor variables, creating a tree structure where each node represents a decision based on a predictor variable. This approach identifies patterns and influential features in the dataset, offering interpretable insights for predicting customer churn in this study. The first step was looking at the accuracy at k = 1, which ended up being 78.85% accuracy at predicting churn. The model was then run with all k values between one and 50 to determine the value with the best accuracy. The best\_k value was determined to be 39 with an accuracy of 84.15% to optimize the model's performance.



## Provide the Code

# Create the decision tree model

tree\_model <- rpart(ChurnYes ~ ., data = train\_data, method = "class")

# Print the summary of the decision tree model

summary(tree\_model)

# Visualize the decision tree

rpart.plot(tree\_model)

# Summary and Implications

## Explain Accuracy and mean squared error

The accuracy of the prediction model, coupled with a Mean Squared Error (MSE) of 0.119, underscores its effectiveness in classifying instances of churn and predicting their outcomes. While accuracy reflects the proportion of correctly classified churn instances, the MSE quantifies the average squared difference between actual and predicted values. In this context, the low MSE value suggests that the model's predictions closely align with the observed churn outcomes, enhancing its reliability in forecasting churn behavior. These metrics together offer comprehensive insights into the model's performance, indicating its proficiency in accurately predicting churn and providing valuable guidance for decision-making processes.

## Discuss Results

The predictive analysis employing a decision tree model yielded promising results, indicating a strong ability to classify churn instances with an accuracy rate suggestive of its efficacy in identifying potential churners. The low Mean Squared Error (MSE) further underscores the model's robust performance. Key predictors such as bandwidth usage, tenure, and service subscriptions emerged, providing valuable insights into churn behavior. These findings offer actionable implications for businesses, enabling the implementation of targeted retention strategies to mitigate churn and foster customer loyalty. Continuous refinement and monitoring of the model will be crucial for adapting to evolving customer dynamics and maintaining competitiveness in the market.

## Limitations of Analysis

One limitation of the data analysis is the potential for overfitting the decision tree model. Overfitting occurs when the model learns not only the underlying patterns in the training data but also the noise and random fluctuations. This can result in a model that performs well on the training data but poorly on unseen test data. Overfitting is particularly common with decision trees because they can create very complex models that perfectly classify the training data. To mitigate this, techniques such as pruning the tree, setting a maximum tree depth, or using cross-validation to tune hyperparameters can be employed to ensure that the model generalizes well to new data (Hastie, Tibshirani, & Friedman, 2009).

## Recommend a Course of action

Based on the results and implications of the analysis, it is recommended that the organization focuses on improving customer retention strategies, particularly targeting customers with characteristics identified as significant predictors of churn. These strategies could include personalized offers or incentives tailored to specific customer segments, such as those with shorter tenure or higher monthly charges. Additionally, investing in enhancing the quality of service, especially in areas highlighted by the analysis as influential, like streaming services or contract terms, could help mitigate churn rates. Regular monitoring of customer feedback and behavior patterns, along with ongoing analysis to identify changing trends or emerging predictors of churn, should also be incorporated into the organization's retention efforts to ensure continued effectiveness and adaptability to evolving customer needs and preferences.

# Code References

Data Camp. (2024). Data Mining I. Retrieved May, 2024, from https://www.datacamp.com

# References

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.