D209 Performance Assessment

Data Mining I

Task 1 Classification 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 monthly charges and tenure of customers who churn compare to those who remain with the telecommunications company, and can K-nearest Neighbor analysis be utilized to predict customer churn based on these attributes within a dataset of 10,000 customers? The accuracy of the prediction will also be quantified.

## Goal of Analysis

The goal is to analyze monthly charges and tenure differences between churned and retained customers and to develop a predictive model using K-nearest Neighbor analysis to forecast customer churn within a dataset of 10,000 customers. This analysis is to understand how different customer segments vary in terms of their churn rates, monthly charges, tenure, and other relevant attributes. This analysis could help the telecommunications company in devising targeted marketing strategies, improving customer retention efforts, and optimizing pricing plans to better serve its customers.

# Method Justification

## Benefits of using KNN

The K-nearest Neighbor (KNN) algorithm is employed to analyze the dataset, comprising attributes like monthly charges and tenure, for classification tasks. In the training phase, it stores all available data points and their corresponding labels. During testing, it calculates the distances between new data points and existing ones, selecting 'k' nearest neighbors. The majority voting based on the labels of these neighbors then determines the predicted label for the new data point. Expected outcomes include the identification of patterns and similarities in the dataset and the prediction of customer churn based on these attributes, facilitating targeted customer retention efforts and decision-making for the telecommunications company.

## Summarize Assumptions

One assumption of the chosen classification method, K-nearest Neighbor (KNN), is that similar data points tend to belong to the same class or category. This assumption implies that data points close to each other in the feature space are likely to have similar characteristics and thus similar labels. Therefore, KNN assumes that proximity in the feature space correlates with similarity in the target variable.

## 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. **caret**: The **caret** package offers a unified interface for training and evaluating machine learning models. It supports the analysis by providing functions for data splitting, model training, hyperparameter tuning, and performance evaluation, streamlining the process of building and assessing predictive models.
5. **class**: This library provides functions for K-nearest neighbor (KNN) classification, a common machine learning algorithm used in predictive modeling. It supports the analysis by enabling the implementation of KNN classification for predicting customer churn based on relevant features in the dataset.
6. **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.
7. **pROC**: The **pROC** package provides functions for Receiver Operating Characteristic (ROC) analysis and calculation of the Area Under the Curve (AUC) metric. It supports the analysis by enabling the assessment of model performance and discrimination ability, particularly for binary classification tasks like customer churn prediction.

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 key data preprocessing goal for the classification method, especially K-nearest Neighbor (KNN), is to scale or normalize the features in the dataset. Since KNN relies on distance measures, ensuring that all features are on a similar scale is crucial. If features have different scales, those with larger magnitudes may disproportionately influence the distance calculation, potentially biasing the results. Techniques such as min-max scaling or standardization can be employed to achieve this goal, ensuring that each feature contributes equally to the distance calculation and enhancing the performance of the KNN algorithm.

## Variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Type** | **Feature or Target** | **Description** |
| Churn | Categorical | Target | Whether the customer has discontinued service within the last month |
| Tenure | Numerical | Feature | The contract term of the customer |
| Monthly Charge | Numerical | Feature | The amount charged monthly |

## Describe cleaning steps and goals

The goal of cleaning the data is to ensure the usability of the data set. The first step is to make sure there is no duplicate data which is accomplished by checking for distinct rows of data. The next step is to check if the data has missing data that needs to be filled in or removed. The data will then be condensed into just the columns needed for this assignment. The Churn column will then be turned into numeric values. The continuous variables will finally be standardized using the z score to ensure they are on a similar scale, which is an important step when analyzing for KNN.

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

#Checking for repeated data

CC <- distinct(CC)

#Checking for missing values throughout dataset

vis\_miss(CC)

# Select columns of interest

SC <- select(CC, Churn, MonthlyCharge,Tenure)

#Inspecting Data

str(SC)

summary(SC)

# Convert "yes" to 1 and "no" to 0

ifelse(SC$Churn == "Yes", 1, 0)

# Standardization (Z-score) - Standardizing only the continuous variables

SC$z\_MonthlyCharge <- (SC$MonthlyCharge - mean(SC$MonthlyCharge)) / sd(SC$MonthlyCharge)

SC$z\_Tenure <- (SC$Tenure - mean(SC$Tenure)) / sd(SC$Tenure)

## 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/D209A/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”

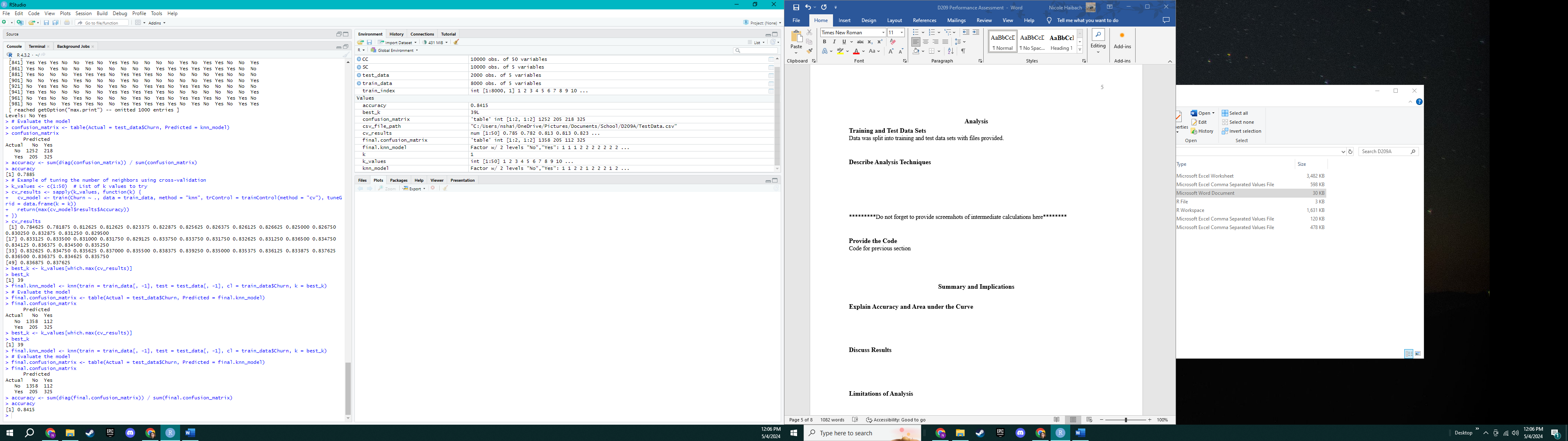
# Analysis

## Training and Test Data Sets

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

## Describe Analysis Techniques

The 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 models performance.



## Provide the Code

# Train the KNN model

k <- 1 # Number of neighbors

knn\_model <- knn(train = train\_data[, -1], test = test\_data[, -1], cl = train\_data$Churn, k = k)

knn\_model

# Evaluate the model

confusion\_matrix <- table(Actual = test\_data$Churn, Predicted = knn\_model)

confusion\_matrix

accuracy <- sum(diag(confusion\_matrix)) / sum(confusion\_matrix)

accuracy

# Example of tuning the number of neighbors using cross-validation

k\_values <- c(1:50) # List of k values to try

cv\_results <- sapply(k\_values, function(k) {

cv\_model <- train(Churn ~ ., data = train\_data, method = "knn", trControl = trainControl(method = "cv"), tuneGrid = data.frame(k = k))

return(max(cv\_model$results$Accuracy))

})

cv\_results

best\_k <- k\_values[which.max(cv\_results)]

best\_k

final.knn\_model <- knn(train = train\_data[, -1], test = test\_data[, -1], cl = train\_data$Churn, k = best\_k)

# Evaluate the model

final.confusion\_matrix <- table(Actual = test\_data$Churn, Predicted = final.knn\_model)

final.confusion\_matrix

accuracy <- sum(diag(final.confusion\_matrix)) / sum(final.confusion\_matrix)

accuracy

# Summary and Implications

## Explain Accuracy and Area under the Curve

The K-nearest neighbor (KNN) classification model achieved an accuracy of 0.8415, indicating it correctly classified approximately 84.15% of instances. The Area Under the Curve (AUC) value, calculated at 0.7426, suggests moderate discriminatory power in distinguishing between positive and negative instances. Accuracy provides a measure of correctness, while AUC offers insights into the model's ability to rank instances and make reliable predictions, particularly for classifying positive and negative cases. These metrics collectively inform our understanding of the model's performance, guiding potential improvements in future iterations.

## Discuss Results

The classification analysis revealed an accuracy of 0.8415 and an Area Under the Curve (AUC) value of 0.7426 for our K-nearest neighbor (KNN) model. These metrics suggest that the model effectively predicts customer churn to a reasonable extent. While the high accuracy indicates good overall performance, the moderate AUC value suggests room for improvement in classifying positive churn cases accurately. These findings offer valuable insights for our telecommunications company. They can guide proactive strategies to retain at-risk customers based on identified predictors like MonthlyCharge and Tenure. However, further refinement of the model may be necessary to optimize its predictive capabilities for real-world applications. Overall, the analysis underscores the potential benefits of employing machine learning in telecommunications for improved customer management and business outcomes.

## Limitations of Analysis

One limitation of our data analysis is the potential oversight of important variables influencing customer churn. While our model considers factors like MonthlyCharge and Tenure, other significant aspects such as customer demographics or service quality might not be included in the dataset (Wang & Feng, 2017). Neglecting these factors could lead to incomplete or biased predictions, impacting the accuracy and applicability of our findings. To mitigate this limitation, future analyses could explore more extensive datasets or employ advanced techniques to capture the full complexity of customer churn dynamics.

## Recommend a Course of action

Based on the analysis results, the organization should implement targeted retention strategies to address customer churn. Utilizing the K-nearest neighbor (KNN) classification model's predictions, the organization can identify customers at risk of churning and tailor retention efforts accordingly. Insights from the analysis, such as significant predictors like MonthlyCharge and Tenure, can inform personalized incentives or service adjustments aimed at retaining at-risk customers. Additionally, further research to identify and incorporate additional variables influencing churn behavior will enhance the model's predictive accuracy. By implementing these strategies, the organization can reduce churn, improve customer satisfaction, and drive business growth.

# Code References

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

# References

Wang, Y., & Feng, H. (2017). Customer churn prediction in telecommunications. Expert Systems with Applications, 73, 82-89. https://doi.org/10.1016/j.eswa.2016.12.008