D209 – Data Mining (Task 2)

Morrell J. Parrish

Western Governors University

Table of Contents

[A. Research Question 3](#_Toc110790139)

[A2. Analysis Goal 3](#_Toc110790140)

[B. Chosen Technique 3](#_Toc110790141)

[B2. Assumption 3](#_Toc110790142)

[B3. Packages/Libraries Use Justification 3](#_Toc110790143)

[C. Data Preparation Description 5](#_Toc110790144)

[C2. Variable Identification and Classification 5](#_Toc110790145)

[C3. Data Preparation Steps 6](#_Toc110790146)

[C4. Cleaned Data Set 7](#_Toc110790147)

[D. Analysis 8](#_Toc110790148)

[D2. Analytical Technique Description 8](#_Toc110790149)

[D3. Classification Analysis Code 12](#_Toc110790150)

[E. Summary 13](#_Toc110790151)

[E2. Classification Analysis Results and Implications 13](#_Toc110790152)

[E3. Analysis Limitation 13](#_Toc110790153)

[E4. Recommended course of Action 14](#_Toc110790154)

[G. Panopto video recording 14](#_Toc110790155)

[References 15](#_Toc110790156)

D209 – Data Mining (Task 2)

# A. Research Question

During this course of research, we will determine which customers are at a higher risk of churn and which features (variables) can be an indicator for churn.

## A2. Analysis Goal

The objective of this analysis is to use the decision tree methodology to determine which variables determine if a customer will churn or not; this analysis will reduce the number of our predictor variables down to the most significant one(s). “The churn rate, also known as the rate of attrition or customer churn; is the frequency in which consumers discontinue doing business with a company. It is commonly represented as the percentage of service subscribers who cancel their memberships within a specified time frame” (Frankenfield, 2022).

# B. Chosen Technique

This analysis will employ the ***decision tree methodology***. The ***decision tree methodology*** is a widely used data mining method for developing prediction algorithms for a target variable or establishing classification systems based on multiple covariates. This model was chosen due to its simplicity; a decision tree is a “flowchart-like tree structure in which each internal node represents a test on an attribute, each branch represents a test outcome, and each leaf node (terminal node) holds a class label” (Geeks for Geeks, 2002).

## B2. Assumption

One assumption of decision trees is that feature values should be categorical. If the values are continuous, they are discretized before the model is built. Recursively, records are distributed based on attribute values.

## B3. Packages/Libraries Use Justification

The following packages/libraries will be used for this analysis:

* Pandas
  + used to read and manipulate data via series (one-dimensional structure) or dataframes (multi-dimensional data structure)
* NumPy
  + used to perform mathematical computations
* Matplotlib
  + used to create visualization (plotting and graphing)
* Seaborn
  + used to create visualization (plotting and graphing)
* Scikit-learn
  + used to perform scientific computations
  + used to split our data into training and test sets
  + used for predicting and classification analysis
* SciPy
  + used for scientific and technical computation
* Graphviz
  + used to create graph objects, which can be completed using different nodes and edges
* DMBA
  + used in data mining for business analytics
* PIL
  + Python imaging library, that adds image processing capabilities

# C. Data Preparation Description

One goal of data preprocessing (**data cleaning/mining**) is to make the training/testing process easier by appropriately transforming and scaling the entire dataset. Before training machine learning models, preprocessing is required. Outliers are removed during preprocessing, and the features are scaled to an equivalent range (Misra et al., 2020).

To use the churn dataset in our analysis we will first need to prepare the data.

The following steps were taken to prepare the dataset for analysis:

* download the churn dataset
* determine which variables will be used in the analysis
* import the dataset into *PyCharm*
* remove independent variables, demographics, and personal identification variables not being used in the analysis
  + caseorder, customer\_id, interaction, UID, city, state, county, zip, lat, lng, population, timezone, job, email, contacts
* determine if any outliners exist and remove them

## C2. Variable Identification and Classification

The **continuous variables** (16)that will be used in this analysis will include age, children, income**,** outage\_sec\_perweek, yearly\_equip\_failure, tenure, monthlycharge, **b**andwidth\_GB\_Year, item1 (timelyresponse), item2 (fixes), item3 (replacements), item4 (reliability), item5 (options), item6 (respectfulness), item7 (courteous), and item8 (listening).The **categorical variables** (19)that will be used in this analysis will include area, marital, gender, churn, techie, contract, portmodem, tablet, internetservice, phone, multiple, onlinesecurity, onlinebackup, deviceprotection, techsupport, streamingtv, streamingmovies, paperlessbilling, paymentmethod.

## C3. Data Preparation Steps

To use the churn dataset in our analysis we will first need to prepare the data:

* import the dataset into *Python (PyCharm)*
* view the dataframe’s description, structure, and data types
* view summary statistics
* evaluate the dataset, remove null or missing values
* remove any outliners
* remove demographics, and personal identification
  + caseorder, customer\_id, interaction, UID, city, state, county, zip, lat, lng, population, area, timezone, job, email, contacts

**The below code was used to prepare our data**:



Text

Description automatically generated

## C4. Cleaned Data Set

The prepared dataset used for this analysis has been uploaded with the assessment file.

# D. Analysis

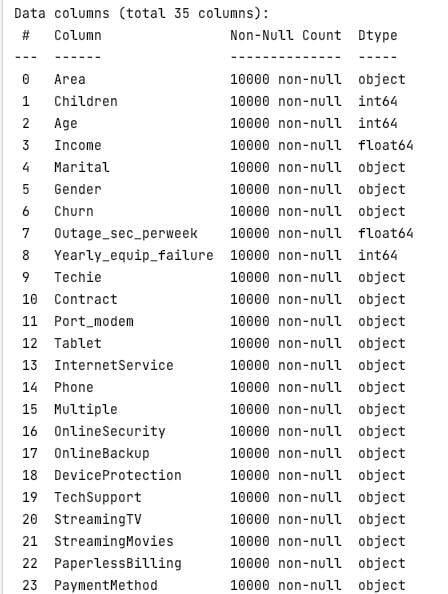
The training and test datasets used for this analysis have been uploaded with the assessment file.

## D2. Analytical Technique Description

Our analytical technique includes the following steps: (1) read in or load the data using Pandas’ ***read( )*** function – in this case it will be our cleaned churn data set, (2) check and verify datatypes using Panda’s ***info ( )*** function, (3) verify summary statistics using the ***describe ( )*** function, (4) set predictor and target variables; split up the dataset into inputs (X) and our target variable (y) using Pandas’ ***drop( )*** function – this allows you to drop the target variable from the dataframe and store it in the variable ‘X’, (5) define both the categorical and numerical features, (6)split the dataset into training and test sets using Scikit-learn’s function ‘***train\_test\_split’,*** (7) extract both training and test datasets, (8) create model - this can be done using the ***DecisionTreeClassifier ( )*** and ***fit*** ***( )*** functions , (9) plot the decision tree variable and boundaries, (10) verify model’s accuracy via the ***classification summary ( )*** function,







Table

Description automatically generated



Table, Excel

Description automatically generated

**Text

Description automatically generated with medium confidence**

**Graphical user interface, text, application

Description automatically generated**

**Text

Description automatically generated**

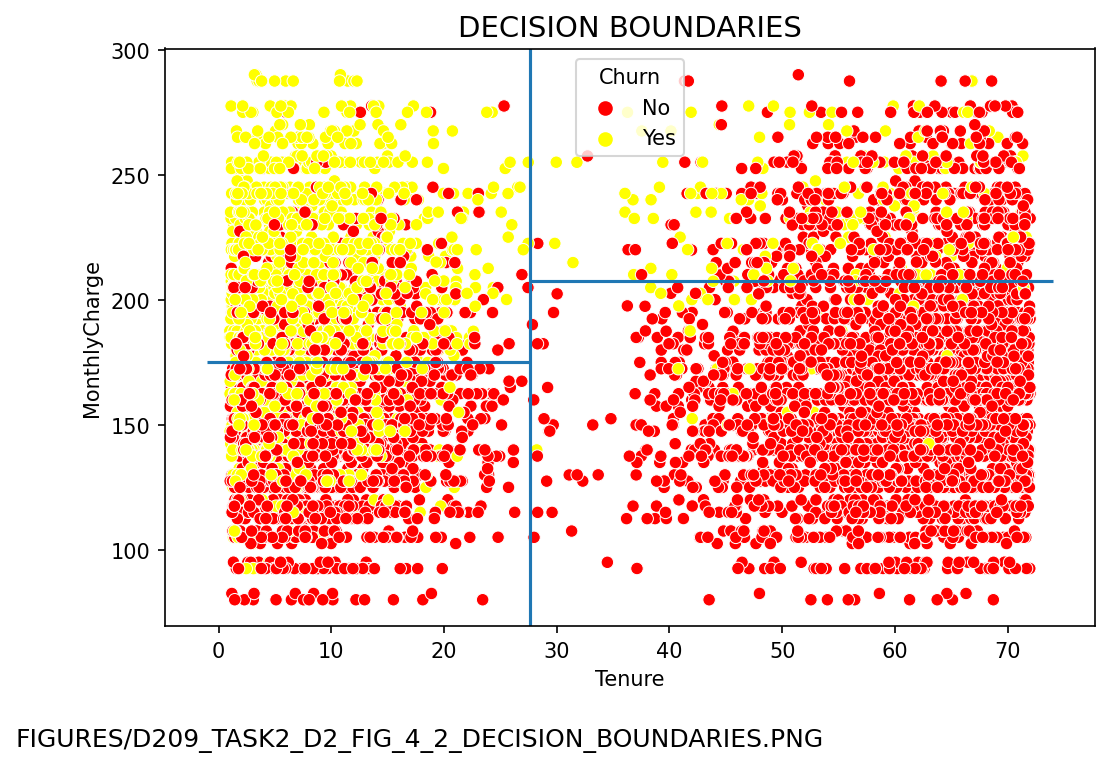
****

**Diagram

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

****

The above scatterplot represents the four terminal nodes of our decision tree

## D3. Classification Analysis Code

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated with low confidence

Text

Description automatically generated

A picture containing text

Description automatically generated

# E. Summary

The training model's classification tree achieved an accuracy rate of 0.8330, or 83**%**; the test model achieved an accuracy rate of 0.8405, or 84**%**. Both models' accuracy was calculated by adding the TP and TN and then dividing the total number of records (TP + TN + FP + FN).

## E2. Classification Analysis Results and Implications

Using tenure and monthly charge as predictor variables, the test model predicted our target class with an 84% accuracy rate; it should also be noted that the predicted class has a 16% chance of being incorrect.

## E3. Analysis Limitation

One limitation of this analysis is that a small change within the dataset can cause variance and make the tree structure unstable; for instance, increasing or decreasing a customer’s monthly charge can affect the outcome.

## E4. Recommended course of Action

I would recommend that the company focus on those individuals where their monthly charge <= $175; those customers tend to churn more. They should look at contracts as well; those customers with a monthly charge <= $175 and with a contract <= 5 months will churn. Focusing on those customers meeting these parameters or criteria will allow the company to provide or create incentives for them to stay; customers with longer tenure tend not to churn – the ultimate goal is to retain as many customers as possible and to increase their tenure.

# G. Panopto video recording

[VideoLink](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=dc40ddca-86be-444c-972b-aeeb0023dc1e)

References

*Decision Tree*. Geeks for Geeks. (2022, July 20). Retrieved August 7, 2022, from <https://www.geeksforgeeks.org/decision-tree/>

Misra, S., Li, H., & He, J. (2020). *Data preprocessing*. Data Preprocessing - an overview | ScienceDirect Topics. Retrieved August 7, 2022, from https://www.sciencedirect.com/topics/engineering/data-preprocessing

WGU. (n.d.). NVM2 TASK 2: *Predictive Analysis*. WGU Performance Assessment. Retrieved July 23, 2022, from https://tasks.wgu.edu/student/000194226/course/20900018/task/2807/overview