TJ Tracy

WA #4.1 Decision Trees

BUS 380

3/27/23

**Context**

The project aims to identify possible dissatisfied customers to be contacted to prevent churn and identify causes of their dissatisfaction. There are two data sets available. The first data set is a training data set that includes a customer’s gender, age, education, duration of service use, and frequency of service use. It also includes a multitude of attributes that attempt to quantify the service’s quality such as Information Accuracy, System Reliability, System Security, Service Competence, etc. There is an ID column and a satisfaction column for each respondent. Satisfaction ranges from being “dissatisfied”, ”neutral”, “satisfied”, “somehow distatisfied”, or “somehow satisfied”. The scoring data set has all the attributes of the training data set but does not include the satisfaction attribute.

**Data Preparation**

In data preparation, the training data’s satisfaction column is changed. All entries of “somehow distatisfied” are replaced with “somewhat dissatisfied” and all entries of “somehow satisfied” are replaced with “somewhat satisfied” to correct the spelling and phrasing of the entries. The attribute “Satisfation” is misspelled and is corrected to “Satisfaction”. The data contains no missing values. Missing values are checked by applying the is.na function to every attribute in the training data and finding the sum, if any, of those missing values. The data pertaining to the service’s quality does contain outliers and those outliers are replaced with the mean of their respective columns to eliminate skew in our dataset and make the results have higher reliability and accuracy. The outliers are found through IQR testing and applying the IQR test to each entry. The ID column is removed on the training data set because ID does not influence a customer’s satisfaction with the service. Therefore, should be removed from the neural network model. The “Satisfaction” column is a character class; however, neural networks in R have difficulty with improperly formatted characters in data. So, the “Satisfaction” attribute is converted to a Factor so the neural network can run on the training data.

**Building the Model**

The project uses a neural network to classify each respondent to their likely satisfaction with the service. Neural networks are flexible and can operate on a wide range of data. Additionally, their outputs are accurate and easy to interpret; however, it is very difficult to understand how the neural network arrived at its decision. Given that information the model is trained on every attribute in the training data set, excluding the ID column.

**Applying the Model**

With neural network, a classification can be made for each consumer in the scoring data set. The classification of the neural network is based on the training data set. The model attaches weights to each attribute and outputs classifications for each response in the scoring data set. The resulting list of consumer classifications is filtered for dissatisfied customers. Therefore, the project's objective is accomplished, and a list of dissatisfied consumers is available to be contacted by the consumer relationship department.

**Supporting Code**

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application

Description automatically generated

**Supporting Visualization(s)**

Text

Description automatically generated

\*Head of list to review

-------------------------------------------------------------------------------------------------------------------------------