**Data Mining and Business Analytics Final Project**

Janna Schlageter

We will gain insight about the similarities between customers and between data points through our data mining processes. We will create visualizations to better understand the data and interpret customer satisfaction and we will develop models to predict customer satisfaction.

This dataset is a record of sales of medical devices for a business. This data includes 6 variables which are, the customer ID, the year of business, the number of sales calls, the number of target emails, the number of sales, and the rating of customer satisfaction. The customer ID is an integer starting at 1 which assigns a label to each individual customer. In this dataset there are 47 individual customers. The “year” variable is an integer between 2009 and 2021 representing the year in which this row of business data with that customer occurred. The number of sales calls is an integer between 0 and 5 indicating the number of sales calls made to that customer in that year. Similarly, the number of targeted emails is an integer between 0 and 3 quantifying the number of emails sent to that customer in the indicated year. The number of sales variable is an integer that indicates the number of sales made that year to that customer and is between 0 and 4. Finally, the “CustomerSatisfaction” column indicates how satisfied the customer was that year with the business. This value can be a -1, a 0, or a 1. The grain of this dataset is the unique customer ID with the year, this can identify each row individually.

In order to prepare this data for data mining techniques, we will begin by exploring the data. We did this with the “.describe()” and “.dtypes” functions. Then used “.value\_counts()” to explore the customer satisfaction column to determine all possible integer values for this variable. We can see that the “year” column is the only variable that isn’t capitalized so for format sake, we will rename this column to “Year”. This completes our data exploration and we may now begin data preparation. Since we will start with predicting customer satisfaction, we will create a list of predictor variables and the outcome variable. We will normalize our data.

The first step in understanding this data and providing some insight to the business owner is to include some data visualizations. We do this by including a bar chart of average customer satisfaction over the years by customer and a histogram of customer satisfaction as well. In order to visualize if there were strong correlations between customer satisfaction and the other variables, we then created a heatmap for the dataset, but did not see any strong correlations (greater than 0.7) between variables.

The first model implemented to investigate the data was a classification tree. We implemented this with the original dataset, and changed the CustomerSatisfaction to a categorical variable. We partitioned the data into a training set and a validation set. For the decision tree we used a max depth of 8, minimum sample split of 5, and a minimum impurity decrease of 0.00001. This gave us an accuracy of 0.8361 on the training data and an accuracy of 0.7265 on the validation data.

The next model we implemented was K-means Clusters. We used this method to determine if certain instances of data could be grouped and provide insight. After several rounds of experimenting with several different k values, it was determined that the clusters are not very tight with values ranging from 44.92 to 207.31.

In conclusion, there are not strong correlations between the different variables in the dataset nor are there tight clusters of customers. Our visualizations provided us with the insight that many customers over the years have an indifferent or negative satisfaction rate when dealing with this business. Our decision tree was fairly accurate, predicting customer satisfaction with an accuracy rate of about 73% on test data. This indicates that we may need to collect additional data to properly predict customer satisfaction or to cluster customers in groups