# D209 Data Mining I

Professor Keiona Middleton

Mackenzie Simon

# Part 1 Research Question:

A1: Our goal when looking at the Telecommunications Churn Data set is to predict customer churn. We will focus on our goal by creating a decision tree algorithm when targeting the dependent variable Churn. After using decision tree learning, we can then use our input parameters to predict whether a customer will churn. After running our decision tree model, we can answer the question “what to invest in to minimize customer churn?”.

A2: The goal of our analysis is to build a model using decision trees for predicting customer churn. Once we build our decision tree model, we can test to see if our model is valid. If our model is valid, we know what variables will help us in predicting churn.

# Part 2 Method Justification

B1: A decision tree algorithm works by building a classification model in the form of a tree. It breaks down each data category into a series of smaller subsets as the tree is developed. The result is a tree with decision nodes and leaf nodes. A decision node consists of two or more branches. Leaf nodes represent a classification or decision. Our decision tree algorithm builds our classification model based on the training data. After we have the training model, we compare the results to our test data. We expect our decision tree to accurately predict whether a customer will churn.

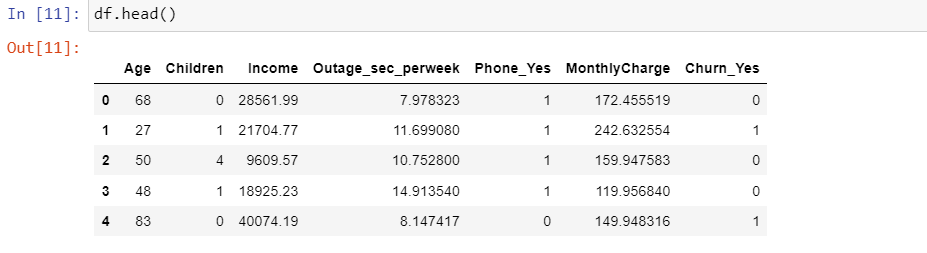
B2: Decision trees are a non-parametric based algorithm and do not require any assumptions on data distribution. Sehra (2018) explains how one of the assumptions for our Decision Tree algorithm is the deeper the tree, the more complex the decision rules.

B3: Some of the benefits of using Python are the libraries created for data analytics. In this project we used NumPy for indexing and arrays, Pandas for data formatting, and SciKit for data modeling. With SciKit we scale our data, construct a training/test model, build a decision tree algorithm, create a classification report, and generate a confusion matrix. Zhidkov, R (2021) described Python as open source and having a large community for problem solving. Python is useful when analyzing large datasets due to its speed and processing. For this project, Python is useful because our data has 10,000 rows and 50 columns.

# Part 3 Data Preparation:

C1: One important pre-classification step for our decision tree model is scaling our data. Data scaling can be achieved by normalizing or standardizing input and output variables. In addition, data scaling allows us to optimize our decision tree processing speed. We use the Scikit Learn preprocessing library for data scaling.

C2: Our variables for our decision tree model are listed below. We will use Age (continuous), Income (continuous), Outage Seconds Per Week (continuous), Phone Plan Yes (categorical), Monthly Charge (Continuous) as our X or predictor variables. We will use Churn (categorical) as our y or target variable.



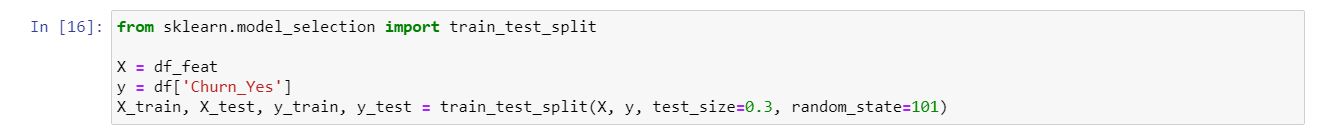
C3:

For my data preparation, I used the Pandas library to import the csv data. I started by converting all the binary columns into dummy variables. I then looked through the data and chose what columns I thought would be useful for predicting churn. I made sure these variables were continuous or binary and would be useful in predicting churn. I then dropped the columns I deemed not useful to our model. I proceeded to standardize our data using the preprocessing package from sklearn. We had to create a scaler model and fit the model to our cleaned data. Once the data was fit to the scaler, we transformed the data. I created two variables, one that scaled all the data under cleaned\_csv and another variable df\_feat which excluded the target variable Churn. I set our transformed data to the variable cleaned\_data in order to print the dataset.



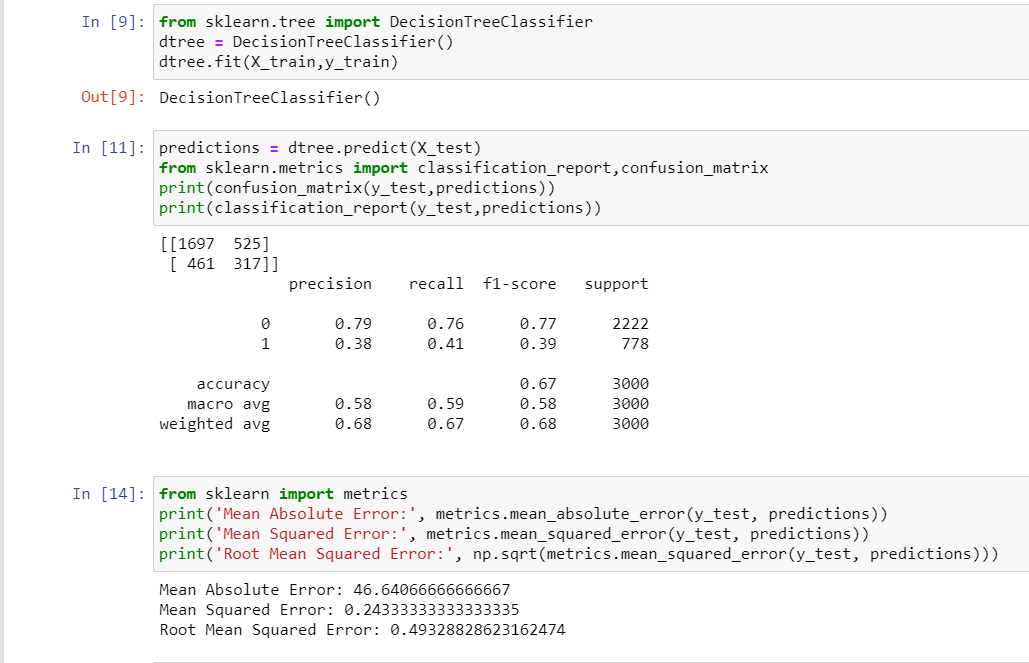
# Part 4 Analysis:

D1: We can use the train\_test\_split function from Sci Kit to setup the testing and training data. We set our scaled data as the independent variables and Churn as the dependent variable.



D2/D3:

In order to setup our decision tree model, we import the DecisionTreeClassifier from sklearn. We then create a decision tree classifier model and set the model to the variable dtree. We then fit our dtree model with the data X\_train and y\_train. Once we have fit our model to the training data, we can analyze our model using the test data. We then create a confusion matrix and classification report to analyze our model. Finally, we import the metrics library from sklearn to analyze our mean squared error.



# Part 5 Data Summary and Implications:

E1. From our classification report, the macro average for our model was .58 for precision, .59 for recall, and .58 for f-1 score. We had an overall f-1 score accuracy of .67. The results show our model isn’t very accurate. From our metrics report, the mean squared error sits at .2433 which also isn’t very strong. I would prefer to lower the mean squared error closer to zero in order to optimize our model.

E2: The results from our analysis show that we need to continue improving our decision tree model. We need to either swap our input variables, continue adding variables to our model, or switch to a different classification model such as random forest.

E3: One of the limitations running a decision tree model is the model can be sensitive to small permutations in the data. A slight change can result in a drastically different decision tree. Decision tree models can also run into errors when dealing with multicollinearity. The decision tree model can highlight one classification method over the other.

E4: Based on our results, I would recommend continuing to improve our model. I would not feel confident in using our model as a predictor for churn. One suggestion I would give the company would be to change our model to a random forest tree. Even though our model is somewhat weak sitting at .67 for f-1 score accuracy, I would suggest continuing to invest in our predictor variables.

# References

Sehra, C. (2020, November 30). *Decision Trees Explained Easily*. Medium. https://chiragsehra42.medium.com/decision-trees-explained-easily-28f23241248.

Zhidkov, R. (2021, January 10). *Why Python is Essential for Data Analysis*. RTInsights. https://www.rtinsights.com/why-python-is-essential-for-data-analysis/#:~:text=The%20object%2Doriented%20programming%20language,streamline%20large%20complex%20data%20sets.&text=Being%20fast%2C%20Python%20jibes%20well,not%20limited%20to%20scientific%20computing.

**Resources for Python Libraries:**

https://matplotlib.org/

https://numpy.org/

<https://pandas.pydata.org/>

https://scikit-learn.org/stable/

https://seaborn.pydata.org/