Part I: Research Question

A1. Proposal of Question

Using Decision Tree regression, is it possible to predict customers' data usage based on the given dataset and identify which variables play a significant role to hopefully understand pathways to improve our overall data infrastructure?

A2. Defined Goal

The goal and objective of this analysis are to use Decision Tree regression to help us understand customers' bandwidth usage and give us an idea of which variable can predict bandwidth usage. By understanding customer bandwidth usage, we can reduce cost, and increase our data infrastructure efficiency. For our analysis, we will be using the variable "Bandwidth_GB_Year" as our response variable and several predictor variables such as "Employment", "InternetService", "Area", etc. to predict customer bandwidth usage.

Part II: Method Justification

B1. Explanation of Prediction Method

The use of Decision Tree regression will allow us to analyze our continuous variable through means of splitting the parent node (our trained dataset) into decision nodes (features of the dataset), and then finally to the leaf node (outcome) (*Decision Trees in Python*, 2022). Another explanation of this is that given a particular data point, the model will run it through the entire decision tree answering True/False until it arrives at a particular outcome. The final prediction will be the average value of the dependent variable in that leaf node (K, 2021). For our analysis, the expected outcome is that our model will accurately predict a particular data point and arrive at the most optimal leaf node (or outcome) of the decision tree regression model.

B2. Summary of Method Assumption

One of the assumptions of the Decision Tree is that "in the initial stage, the whole training data should be considered as the root." (Patel, 2022). A summarization of this is that we want to be sure that our trained dataset is the sample that our model is being iterated from so that we can later test our model with test data (Richer, 2021).

B3. Packages or Libraries list

For this analysis, we will be using Python as it has a wide array of libraries and packages that is suited for our purposes. The packages and libraries necessary are numpy, pandas, scikit-learn, Graphviz, and seaborne among other libraries. The versatility of Python and its relationship to statistics and machine learning also make it a perfect fit for our use in this analysis (Talab, Z., 2021). Python will allow us to prepare our data during the data cleaning process using numpy and pandas. We can then use it to further plot our data during the exploratory analysis using libraries like seaborne to visualize our predictor and response variables to detect outliers. Finally, Python with its diverse libraries will allow us to perform our necessary predictive analysis during our data modeling phase using scikit-learn and Graphviz for model visualization. A list of packages or libraries used for this analysis was also provided below:

- Pandas
- Numpy
- Scipy
- Matplotlib
- Seaborn
- Sklearn
- Graphviz

Part III: Data Preparation

C1. Data Preprocessing

One data preprocessing goal is to encode all categorical data that has more than two levels into dummy variables using k-1. Even though decision trees can perform with categorical data in their natural form, dummy encoding has been shown to improve model performance (Johnson, 2019). This will ensure that we have the best model possible.

C2. Data Set Variables

In our raw data, we have a total of 10,000 observation rows with 52 variable columns. After the data cleaning process, we selected 1 response variable and 22 predictor variables of which we later expanded to 27 after creating dummy variables. Since our research question is to understand and predict a customer's bandwidth usage, our target variable is "Bandwidth_GB_Year". Our predictor variables are "Area", "Employment", "Income", "Contract", "InternetService", "Phone", "Multiple", "OnlineSecurity", "OnlineBackup", "DeviceProtection", "TechSupport", "StreamingTV", "StreamingMovies", "Timely_Fixes", "Timely_Response", "Timely_Replacement", "Reliability", "Options", "Respectable_Response", "Courteous_Exchange", and "Evidence_of_active_listening". Before we do our initial analysis, we will create dummy variables for categorical variable "Area", "Employment, "Contract" and then remove one level to avoid dummy variable trap (*Categorical Predictors: How Many Dummies to Use in Regression vs. k-Nearest Neighbors*, 2015). These new dummy variables will be "Dummy_Urban", "Dummy_Suburban", "Dummy_Part Time", "Dummy_Retired", "Dummy_Student", "Dummy_Unemployed", "Dummy_One year", "Dummy_Two Year". Below is the list of which variables are continuous and categorical:

Categorical Variable:

- Area, Employment, Contract, Dummy_Urban, Dummy_Suburban, Dummy_Part Time, Dummy_Retired, Dummy_Student, Dummy_Unemployed, Dummy_One year, Dummy_Two Year, InternetService, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies.

Continuous Variable:

- Bandwidth_GB_Year, Income, Timely_Fixes, Timely_Response, Timely_Replacement, Reliability, Options, Respectable_Response, Courteous_Exchange, and Evidence_of_active_listening.

C3. Steps for Analysis

To prepare the data for analysis, we will:

1. Import all necessary libraries and datasets into python.
2. Identify the unique entries for each of the variables that we are working with.
3. Drop all unused/duplicate columns, impute missing data/remove them, and remove duplicate rows.
4. Recode all binary categorical columns to 1 and 0.
5. Recode categorical columns with more than two levels to an ordinal variable 0, 1, 2, for use on later analysis. These will be later recoded to dummy variables.
6. Update the name of all the Customer Survey columns using the key names provided by the course document (1=Timely_Response, 2=Timely_Fixes, 3=Timely_Replacement, etc)
7. Find outliers by creating a Z-Score or identifying them through univariate/bivariate analysis.

8. Create dummy variables for categorical data that have more than two levels for our Decision Tree regression model.

Part IV: Data Summary and Implications

E1. Accuracy and MSE

Since the model is regression, the accuracy of the model is measured using mean squared error (MSE), mean absolute error (MAE), root mean square error (RMSE), and R-squared (r2) (S., 2021). For MSE, MAE, RMSE, the closer the value is to zero, the better it indicates how well the model fits the data (Z., 2021b). For R-squared, the closer the value is to 1, the greater proportion of variances are accounted for by the model (Bock, 2020). In our analysis, after hyper-tuning our model, we recorded the following value for our accuracy metric: MSE=4203274.89, MAE=1857.67, RMSE=2050.19, r2=0.004.

E2. Results and Implications

The result of our analysis indicates that the model is not a good fit for the data. Before hyper-tuning our data, our accuracy metric was poor (MSE=8648585.59, MAE=2362.22, RMSE=2940.85, r2=-1.05). After hyper-tuning the data with the best fit parameter, our accuracy metric improved marginally (MSE=4203274.89, MAE=1857.67, RMSE=2050.19, r2=0.004.). Through hyper-tuning, we did identify 3 variables that were important for our analysis (“StreamingTV”, “StreamingMovies”, “OnlineBackup”). When compared to previous models such as Multiple Regression, these same variables still produced wildly sporadic accuracy metrics. This indicates that there may be a predictor variable affecting our response variable that was not included in our analysis. With these values, the implications of the results conclude that our Decision Tree Regression model is not a good fit for the data and therefore not a good predictor of a customer’s bandwidth usage.

E3. Limitation

One of the limitations in our analysis is since decision tree regression models are prone to overfitting, we can assume this is what happened here. Our initial data analysis without hyper-tuning produced a poor accuracy metric and had “Income” as its most important variable at .20 with the next variable being “Evidence_of_active_listening” at 0.05. Perhaps we could have performed an initial variable selection but since we had planned to perform hyper-tuning on the model, it would have been irrelevant. Since hyper-tuning did indicate which variables were important, next time we can iterate with just those variables (“StreamingTV”, “StreamingMovies”, “OnlineBackup”). Still, that may not provide us with a conclusive picture as those same variables were found to have a low R-squared when modeled on a Multiple Regression analysis. It seems that there may be some issues within the dataset that is affecting our analysis, or we may have a missing variable.

E4. Course of Action

Though our model performed poorly in accurately predicting customer bandwidth usage (MSE=4203274.89, MAE=1857.67, RMSE=2050.19, r2=0.004.), we were able to identify variables that were important to the overall picture of bandwidth usage. Building from this analysis, we can improve our understanding of customers’ bandwidth usage by reiterating on those discovered features. As those same variables were found to be statistically significant in other models we have worked on, it stands to prove that these variables do play some role in accurately predicting customer bandwidth usage. A recommended course of action is to continue data collection for these variables and to reach out to customers for more information on what other services they are using that may be considered data intensive. To improve customer interactions and data collection, we can offer discounts or incentives to help us further identify the missing link within our data. Our continued goal is to produce a model that

can accurately predict customers' bandwidth usage so that we can improve the efficiency of our data infrastructure.

Part V: Demonstration

F. Panopto Recording

G. Sources for Third-Party Code

Decision Trees in Python. (2022). Engineering Education (EngEd) Program | Section. https://www.section.io/engineering-education/decision-tree-in-python/#:%7E:text=A%20regression%20decision%20tree%20is%20a%20tree%20created,broad%20categories%3B%20Decision%20node%20and%20a%20Leaf%20node.

Kharwal, A. (2021, June 23). *Visualize a Decision Tree in Machine Learning*. Data Science | Machine Learning | Python | C++ | Coding | Programming | JavaScript. https://thecleverprogrammer.com/2020/08/22/visualize-a-decision-tree-in-machine-learning/

Kumar, N., Kumar, N., & Profile, V. M. C. (2022). *Implement Decision Tree Algorithm in Python using Scikit Learn Library for Regression Problem*. Professional Point. https://theprofessionalspoint.blogspot.com/2019/02/implement-decision-tree-algorithm-in_22.html

S. (2018, October 9). *09 - DecisionTree + GridSearchCV*. Kaggle. https://www.kaggle.com/shotashimizu/09-decisiontree-gridsearchcv

H. Sources

Bock, T. (2020, November 23). *8 Tips for Interpreting R-Squared*. Displayr. https://www.displayr.com/8-tips-for-interpreting-r-squared/

Categorical predictors: how many dummies to use in regression vs. k-nearest neighbors. (2015). BzST. http://www.bzst.com/2015/08/categorical-predictors-how-many-dummies.html

K, G. M. (2021, December 15). *Machine Learning Basics: Decision Tree Regression - Towards Data Science*. Medium. https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda

Johnson, M. K. A. K. (2019, June 21). *5.7 Factors versus Dummy Variables in Tree-Based Models | Feature Engineering and Selection: A Practical Approach for Predictive Models*. Bookdown. https://bookdown.org/max/FES/categorical-trees.html

Patel, A. (2022, January 6). *Decision Tree - Akash Patel*. Medium. https://imakash3011.medium.com/decision-tree-91adec4370d1#:%7E:text=%20Following%20are%20some%20assumptions%20of%20the%20decision,is%20required%20for%20the%20categorical%20variables.%20More%20

Richer, V. (2021, December 8). *Understanding Decision Trees (once and for all!)* 🤜. Medium. https://towardsdatascience.com/understanding-decision-trees-once-and-for-all-2d891b1be579

S. (2021, August 10). *What is Mean Squared Error, Mean Absolute Error, Root Mean Squared Error and R Squared?* Studytonight. https://www.studytonight.com/post/what-is-mean-squared-error-mean-absolute-error-root-mean-squared-error-and-r-squared

Talab, Z. (2021, November 12). *Benefits of Python for AI*. Developer.Com. https://www.developer.com/languages/benefits-of-python-for-ai/#:%7E:text=In%20addition%20to%20its%20versatility%2C%20Python%20is%20a,platforms%3B%20Windows%2C%20MacOS%2C%20Linux%2C%20Unix%2C%20and%20many%20more.

Z. (2021b, May 10). *What is Considered a Good RMSE Value?* Statology.

https://www.statology.org/what-is-a-good-rmse/