**Part I: Research Question**

1. Can we estimate if a customer will churn based on their services, demographic details, and survey responses?

2. The goal of this analysis is to produce a Decision Tree to estimate if a customer will churn based on several variables in our data set. This would give a company valuable insight into why customers churn, how to anticipate churn, and how to prevent customers from churning.

**Part II: Method Justification**

1. Decision Trees are a machine learning algorithm that can accept both categorical and numerical data. This makes it a good fit for my data set. It is also not sensitive to outliers which could be good but not necessarily a concern with this data. Decision trees process data by splitting the data into smaller segments based on certain conditions that create branches. These Branches lead to a leaf node which in this case would be churn or no churn. I’d expect this decision tree to estimate customer churn based on demographic details, survey response and services.

2. Decision Trees assume that the variables are not parametric. This assumption means that the variables between themselves and between the target variable are not highly correlated.

3.  List the packages or libraries you have chosen for Python or R, and justify how each item on the list supports the analysis:

* + - Pandas: Managing the dataset
    - Numpy: Performing mathematical operations on arrays.
    - MatPlotlib: Graphing.
    - Seaborn: Graphing.
    - Sklar: Statistical modeling.

**Part III: Data Preparation**

1. The goal of the data preparation phase is to select, transform and train the data in a way that will produce a strong Decision Tree algorithm.

1. **Numeric Features** - These are numeric variables that include quantitative data.
   * **Age**: Numeric (represents the age of the customer).
   * **Income**: Numeric (represents the customer's income).
   * **Children**: Numeric (represents the number of children a customer has).
   * **Outage\_sec\_perweek**: Numeric (represents the number of seconds per week the customer experiences outages).
   * **Email**: Numeric (represents the number of emails sent to the customer).
   * **Contacts**: Numeric (represents the number of contacts between the customer and the company).
   * **Yearly\_equip\_failure**: Numeric (represents the number of equipment failures experienced yearly).
   * **MonthlyCharge**: Numeric (represents the monthly charge billed to the customer).
   * **Bandwidth\_GB\_Year**: Numeric (represents the bandwidth usage in gigabytes per year).
2. **Categorical Features** - These variables contain qualitative data.
   * **Techie**: Categorical (A binary feature indicating whether the customer considers themselves tech-savvy).
   * **Contract**: Categorical (represents the type of contract the customer has, e.g., month-to-month, one year, two years).
   * **Port\_modem**: Categorical (binary, indicating whether the customer has a portable modem).
   * **Tablet**: Categorical (binary, indicating whether the customer owns a tablet).
   * **InternetService**: Categorical (type of internet service, e.g., DSL, Fiber Optic, None).
   * **Phone**: Categorical (binary, indicating whether the customer has a phone service).
   * **Multiple**: Categorical (binary, indicating whether the customer uses multiple lines).
   * **OnlineSecurity**: Categorical (binary, indicating whether the customer has online security services).
   * **OnlineBackup**: Categorical (binary, indicating whether the customer uses online backup services).
   * **DeviceProtection**: Categorical (binary, indicating whether the customer has device protection).
   * **TechSupport**: Categorical (binary, indicating whether the customer has tech support services).
   * **StreamingTV**: Categorical (binary, indicating whether the customer uses streaming TV services).
   * **StreamingMovies**: Categorical (binary, indicating whether the customer uses streaming movie services).
   * **PaperlessBilling**: Categorical (binary, indicating whether the customer has opted for paperless billing).
   * **PaymentMethod**: Categorical (represents the method used by the customer for payments, e.g., electronic check, mailed check, bank transfer, credit card).
3. The first step was to convert binary variable to 1 , 0 for Yes, No values.

A screenshot of a computer program

Description automatically generated

Second, I filtered the data for the variables I wanted to include in the Decision Tree algorithm. I chose a breadth of variables that cover customer demographic, services, usage, and customer engagement.

A screenshot of a computer program

Description automatically generated

Third, I used one hot encoding to convert the variables with dtype object to dummie variables. This is important so they can be properly fed into the algorithm.

A close-up of a computer code

Description automatically generated

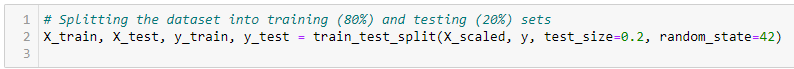
Last, I separated the target variable (‘Churn’) and normalized the remaining numeric data. This is unnecessary to scale the data for decision tree, but I am reusing code from part 1 and it has no impact on the results.

A screenshot of a computer code

Description automatically generated

**Part IV: Analysis**

1. I split the data into training and test data using an 80/20 split as per the instructional video for part 2.



Provide a copy of the cleaned data set. – ***See Attached***

A screenshot of a computer program

Description automatically generated

2. The decision tree model was quite simple to run, and I also ran an accuracy score and classification report to get a quick look at the model's performance.

A screenshot of a computer

Description automatically generated

**Part V: Data Summary and Implications**

1. To summarize, the MSE score is 0.207. The lower this score the better and this is a good score considering the dataset and the model. This indicates the model has a good ability to estimate if a customer will churn given the inputs.

A white box with black text

Description automatically generated

2. As stated above the model has an accuracy score of .793 and MSE score of .207 which means it is accurate at correctly classifying inputs to churn or no churn. When we look at the classification report, we will see that the model is particularly good at classifying non churn instances. This makes sense given the data set and its distribution of churn and no churn. That would also be a limitation of this classification model because it is stronger at predicting actual negatives. There is a normal tradeoff between these two abilities.

1. My recommendation would be to employ this model specifically in the light of estimating if a customer will not churn. Given this information we could build a profile of a loyal customer and aim to onboard or build customer profiles that match that.

**Part VI: Demonstration**

F. Will upload once the rest of the submission has been reviewed and I know I do not have to change the code.

**G. Web Sources.**

<https://www.datacamp.com/tutorial/decision-tree-classification-python>

<https://www.analyticsvidhya.com/blog/2021/04/beginners-guide-to-decision-tree-classification-using-python/>