Part I: Research Question

1. Describe the purpose of this data mining report by doing the following:
   1. Propose one question relevant to a real-world organizational situation that you will answer using one of the following prediction methods:
      1. random forests

A. I will use the random forests and the following predictors to identify which customers are at high risk of churn?

* + - 1. Children
      2. Income
      3. Tenure
      4. Bandwidth\_GB\_Year
      5. Age
  1. Define one goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.

A. The stakeholders can review the data provided by the analysis and create incentives to keep the customers likely to terminate their contracts with the company. This will lead to a lower churn rate.

Part II: Method Justification

B. Explain the reasons for your chosen prediction method from part A1 by doing the following:

1. Explain how the prediction method you chose analyzes the selected data set. Include expected outcomes.
   1. The 'random forest' algorithm improves accuracy by extending on 'single decision tree' where 'multiple decision trees’ are built with random subsets of the data. This extension ensures randomness and grows the decision tree forest. Additionally, it searches for the best feature among the random forest instead of the most important. This ensures there is a wide diversity which leads to a better model. It is one of the most accurate learning algorithms available (Richmond, 2016). I expect the outcome will improve accuracy and decrease overfitting
2. Summarize one assumption of the chosen prediction method.
   1. I assume that the random forests have no formal distributional assumption and are non-parametric. It can handle skewed, multi-modal and categorical data that are ordinal or non-ordinal (Richmond, 2016).
3. List the packages or libraries you have chosen for Python or R, and justify how each item on the list supports the analysis.

A. I will utilize Python due to my previous interaction with it and its Pandas and Sklearn modules. Additionally, I will be using Jupyter notebook as the IDE because it provides a user-friendly experience. Pandas is an excellent package for working with data set as it makes it easy to load and manipulate columns and rows to replace null values. I will use the ‘train\_test\_split’ module to split the data between training and test using the ‘test\_size’ with ‘.20’ parameter to split it 80/20 and ‘random\_state’with ‘21’ for shuffling before splitting. Then RandomForestRegressor will be used to instantiate the regression algorithms. ‘mean\_squared\_error’ and ‘accuracy\_score’ are used to produce the error score and provide accuracy of the model, respectively.

### Part III: Data Preparation

### C. Perform data preparation for the chosen data set by doing the following:

1. Describe one data pre-processing goal relevant to the prediction method from part A1.

A. Select the appropriate independent variables and hyper-parameters for the analysis.

1. Identify the initial data set variables you will use to analyse the prediction question from part A1 and group each variable as continuous or categorical.

A. I will be using the following independent variables to analyze and predict the future bandwidth usage:

##### *Categorical Predictor:*

• Churn

##### *Continuous Predictor:*

• Children • Income • Tenure • Bandwidth\_GB\_Year • Age

1. Explain the steps used to prepare the data for the analysis. Identify the code segment for each step.

1. Use Pandas to import the CSV file in the data frame.

2. Examine and ensure data type consistency in the columns.

3. Identify and resolve spelling mistakes in column headers or row-level data.

4. Validate there are no Null values. If so, remove them.

5. Run regression on the prepared data set.

C.4 Provide a copy of the prepared data set.

Graphical user interface, text, application

Description automatically generated

### Part IV: Analysis

#### D. Perform the data analysis and report on the results by doing the following:

1. Split the data into training and test data sets and provide the file(s).

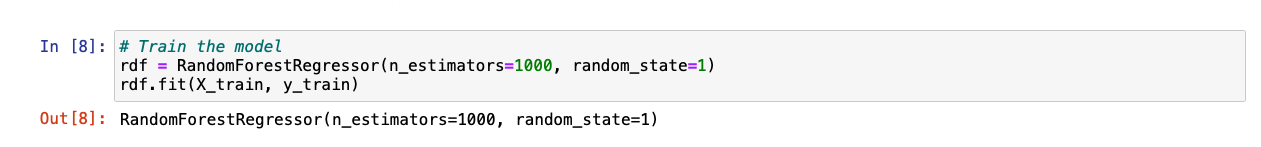
Graphical user interface, text, application

Description automatically generated

1. Describe the analysis technique you used to appropriately analyze the data. Include screenshots of the intermediate calculations you performed.

A. I used the Random Forest algorithm to build several decision trees and then merge them to get an accurate and stable prediction. The algorithm grows the model and adds randomness by searching for the best feature while splitting a node. This will result in diversity that leads to a better model.

The below example shows the algorithm flow. Chart, radar chart

Description automatically generated 

1. Provide the code used to perform the prediction analysis from part D2.



### Part V: Data Summary and Implications

#### E. Summarize your data analysis by doing the following:

1. Explain the accuracy and the mean squared error (MSE) of your prediction model.

Graphical user interface, text, application

Description automatically generated

1. Discuss the results and implications of your prediction analysis.

The churn variable has a low MSE score of .14, suggesting a high accuracy rate at predicting which customer will churn based on the dataset. However, since Mean Square Error is a square of averages, a single high value will lead to a higher mean. This makes MSE vulnerable to large outliers.

The model adequately predicts the Churn rate but could perform better if further analysis is performed on the independent variables.

1. Discuss one limitation of your data analysis.

There are a couple of limitations:

* The algorithm requires more trees to have an accurate prediction, which results in a slower model.
* It doesn't predict beyond the range in the training data, which can lead to overfitting in a noisy dataset. Thompson, B. (2021, December 13)

1. Recommend a course of action for the real-world organizational situation from part A1 based on your results and implications discussed in part E2.

The stakeholders need to analyse which features are common among the customers that have left the company in the past. This will help them identify which customers are at higher risk of churn and provide them with additional relevant services at a competitive price. This will show the customers that the company is a one-stop shop for all their needs, which will lead them to stay with the company for a long time.

### Part VI: Demonstration

#### 1. Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and a summary of the programming environment.

Panapto: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=0d1c8144-b02f-44b4-a2f5-ae2c014ef5e3>

#### G. Record the web sources used to acquire data or segments of third-party code to support the analysis. Ensure the web sources are reliable.

Pandas. (2021). Pandas DataFrames. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.dtypes.html

Get started with references. (2021). Jupyterbook. https://jupyterbook.org/tutorials/references.html#tutorials-references

Marques, A. M. (2020, March 11). How to show all columns / rows of a Pandas Dataframe? Towards Data Science. https://towardsdatascience.com/how-to-show-all-columns-rows-of-a-pandas-dataframe-c49d4507fcf

Starmer, J. (2018, March 5). StatQuest: Logistic Regression. YouTube. https://www.youtube.com/watch?v=yIYKR4sgzI8&t=121s

V. (2019, July 21). Pandas: Apply a function to single or selected columns or rows in Dataframe. ThisPointer. https://thispointer.com/pandas-apply-a-function-to-single-or-selected-columns-or-rows-in-dataframe/

Wijaya, C. Y. (2021, December 15). 5 Must-Know Dimensionality Reduction Techniques via Prince. Medium. https://towardsdatascience.com/5-must-know-dimensionality-reduction-techniques-via-prince-e6ffb27e55d1

Thompson, B. (2021, December 13). A limitation of Random Forest Regression - Towards Data Science. Medium. https://towardsdatascience.com/a-limitation-of-random-forest-regression-db8ed7419e9f

Great Learning Team. (2022, January 17). Mean Squared Error - Explained | What is Mean Square Error? GreatLearning Blog: Free Resources What Matters to Shape Your Career! <https://www.mygreatlearning.com/blog/mean-square-error-explained/>

Richmond, S. (2016, March 21). Algorithms Exposed: Random Forest | BCCVL. Bccvl.Org.Au. Retrieved 10 January 2022, from <https://bccvl.org.au/algorithms-exposed-random-forest/>

Donges, N. (2021, September 17). A Complete Guide to the Random Forest Algorithm. Built In. https://builtin.com/data-science/random-forest-algorithm

#### H. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased or summarized.

Chantal D. Larose, & Daniel T. Larose. (2019). Data Science Using Python and R. Wiley.