**Part I: Research Question**

A.  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 clustering techniques:

•  k-means

I will use the K-means algorithm and the following predictors to identify which customers are at high risk of churn?

* Children
* Income
* Tenure
* Bandwidth\_GB\_Year
* Age
* Churn

2.  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.

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

**Part II: Technique Justification**

B.  Explain the reasons for your chosen clustering technique from part A1 by doing the following:

1.  Explain how the clustering technique you chose analyzes the selected dataset. Include expected outcomes.

I will use KMeans to find the cluster of customers that are likely to churn based on the predictor identified in A1. This will be performed using the Kmeans algorithm to group the data together into clusters based on the distance between points as a measure of similarity using the k average. Each predictor will have one of these points, called ‘Centroid’. The closer a data point (i.e customer) is to a Centroid, it means they will belong to that cluster, and in our situation that will be either Yes or No of the churn column.

Expected outcomes: I expect that by using the KMeans and the predictors identified in A1, the data will be clustered based on their distance to predict their 'Churn' status 'Yes' or 'No'.

2.  Summarize **one** assumption of the clustering technique.

The KMeans algorithm assumes that data is composed of distributions. As the distance from the distribution's center increases, the probability that a point belongs to the distribution decreases. (2019, November 29).

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, matplotlib, and Scipy modules. Additionally, I will be using the Jupyter notebook as the IDE because it provides a user-friendly experience. Pandas is an excellent package for working with data sets as it makes it easy to load, manipulate columns and/or rows to replace null values, and scaled the dataset used by the KMeans.

**Part III: Data Preparation**

C.  Perform data preparation for the chosen dataset by doing the following

1.  Describe **one** data pre-processing goal relevant to the clustering technique from part A1.

* 1. I used the StandardScaler to scale the data as KMeans relies on the distance between the points.

1. Identify the initial dataset variables that you will use to perform the analysis for the clustering question from part A1, and label each as continuous or categorical.

##### The predictors from the A1:

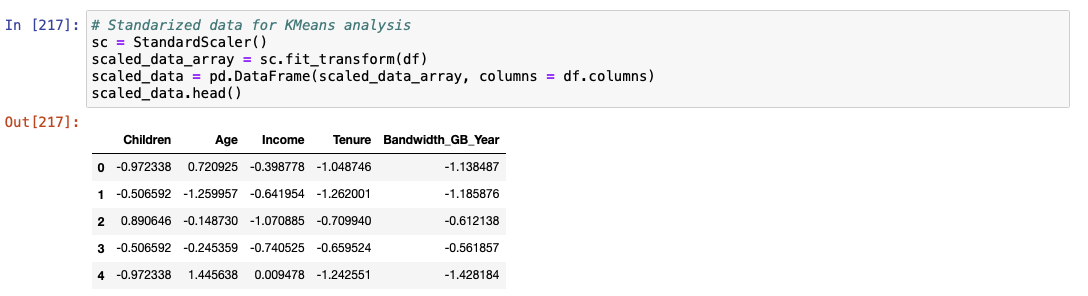
Categorial Predictor:

* Churn

##### Continuous Predictor:

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

1. Explain each of the steps used to prepare the data for the analysis. Identify the code segment for each step.
2. Import dataset to a DataFrame in Pandas.
3. Review the summary statistics of DataFrame, structure & data types.
4. Remove irrelevant columns from the DataFrame "Customer\_id", and "zip code".
5. Text

   Description automatically generated I will normalize the data to bring it into a similar range of values. This will also increase the prediction accuracy as the data point are similarly scaled. It is also important as the KMeans analysis is based on the distance between points.
6. ****Categorical data conversion



1. Provide a copy of the cleaned dataset.
   1. A copy of ‘prepared\_dataset.csv’ has been attached.

**Part IV: Analysis**

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

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

I used the Inertia measure to identify the quality of the clusters. Inertia measures the spread of the cluster Bruce, P., Bruce, A., & Gedeck, P. (2020). A lower value signifies how far the data point is from the centroid of its cluster. In my analysis, the measure was run 10 times to identify the Inertia value of 2, this fits perfectly as we understand that data belong to customer churn records. It is intuitive to try and correlate these 2 clusters with the customer churn, one cluster of customers who churned out, and another for customers who stayed with the company. Next, the churn data were fitted to the model and its predictions were added to the dataset to measure the accuracy.

2.  Provide the code used to perform the clustering analysis technique from part 2.

Chart

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**Part V: Data Summary and Implications**

E.  Summarize your data analysis by doing the following:

1.  Explain the accuracy of your clustering technique.

Using the metric library, KMean’s overall prediction accuracy was approximated at 70.81% for the churn behavior. The KMeans label has two values 0 or 1. The model was able to cluster 2366 data items under label 1. It is very close to the actual count of 2650 that counted the occurrence of ‘Yes’ under the churn column. This gives us a churn prediction accuracy of 89.28%. Similarly, 4715 were identified to be under cluster 0, which is low compared to 7350 with count ‘No’ in the original dataset. This shows a strong correspondence between cluster 1 and its churn behavior. This helped establish that Cluster 0 corresponds to churn ‘No’ and Cluster 1 to ‘Yes’.

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

KMeans is an unsupervised learning method that was able to predict 2 clusters with no prior knowledge about the data. One cluster corresponded with the customer who left the company and the other cluster that primarily showed the count that stayed. KMean was successfully able to identify without any churn-related data the customers who are likely to churn with 89.28% accuracy. This will help stakeholders prioritize the customers’ that need to be communicated to and provided with incentives to deter them from leaving and decrease the churn rate as established in A1.

1. Discuss **one** limitation of your data analysis.

One of the main limitations is that we do not have enough data points to come to an absolute decision. The company needs to start gathering additional data points which will lead to an accurate picture of the customer profile.

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.

As mentioned in E2, the model was successfully able to predict the accuracy of churn with 89%. This tool can now be used to create a profile of the customer that is likely to leave the telecom company in near future. The stakeholders need to identify these customers and engage them via email, social media, and or phone to provide incentives to keep them with the company.

**Part VI: Demonstration**

F.  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.

Panopto Link:

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.

https://developers.google.com/. (2019, November 29). *Clustering Algorithms | Clustering in Machine Learning |*. Google Developers. Retrieved June 18, 2022, from https://developers.google.com/machine-learning/clustering/clustering-algorithms

Sharma, A. (2022, January 6). *Unsupervised Learning using KMeans Clustering - Data Science on Customer Churn Data*. Medium. https://medium.com/data-science-on-customer-churn-data/unsupervised-learning-using-kmeans-clustering-629b26e1ce1e

GeeksforGeeks. (2018, November 28). *How to select multiple columns in a pandas dataframe*. https://www.geeksforgeeks.org/how-to-select-multiple-columns-in-a-pandas-dataframe/

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

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical statistics for data scientists : 50+ essential concepts using r and python*. O'Reilly Media, Incorporated.