**K-means**

**Full Name**

**(Reg. No)**

Table of Contents

[Introduction. 3](#_Toc142470228)

[Objectives 4](#_Toc142470229)

[Description of Dataset 4](#_Toc142470230)

[Prepossessing 5](#_Toc142470231)

[DM Area and DM Algorithm 8](#_Toc142470232)

[Exploration of the Effect of k in k-means clustering 8](#_Toc142470233)

[Results and Analysis 10](#_Toc142470234)

[Conclusion 14](#_Toc142470235)

[References 15](#_Toc142470236)

# Introduction.

Data mining refers to exploration and extraction of large batch of raw unstructured or structured data with the aim of identifying patterns, trends and extract useful information. The extracted information can be used by companies to predict market trends, increase sales and decrease cost. Predictions are done by training Machine Learning models that are able to efficiently to predict the future trends, prices, and recommend products to customers among other things. Data can be classified in several classes like 2 or even more. When we have raw data that is already know which classes is in, then this is supervised learning. Other data maybe just features that are not in any specific classes and this is unsupervised learning. Market forecasting and customer retention are some of the applications of supervised learning while recommender systems and customer segmentation fall in unsupervised learning category.

K-means was first coined by James MacQueen in 1967. In his article “Some Methods for Classification and Analysis of Multivariate Observations”, J. MacQuuen proposed an algorithm that partitioned an instance into small variance. The algorithm did not go by the name k-means but had several names namely: nearest centroid sorting, dynamic clustering, and h-means among other names. Other scholars such as Hugo Steinhaus (1956) and Stuart Lloyd (1957) had previously proposed algorithms similar to k-means. In his article “Sur la Division des Corps Matérielsen Parties”, Steinhaus proposed an algorithm in which the problem could be partitioned into heterogeneous solid by enough selection of partitions. He was the first researcher to propose an algorithm for multidimensional instances (Pérez-Ortega e.t.l, 2019).

K-means is an unsupervised Machine Learning algorithm that groups unlabeled data into clusters based on how similar or different they are to one another. K-means works in 3 steps as follows (K Means Clustering - Introduction - GeeksforGeeks, 2019):

1. Initialize random k-points, known as cluster centroids or means.
2. Categorizing each item (data point) to the closest means and then updating its coordinate based on the centroid it is close to.
3. Repeat the processes for a number of iterations until we get the clusters.

# Objectives

To use k-means clustering to e.g. cluster company dataset in order to help the company understand customer churn and identify patterns, factors, and indicators that contribute to customer attrition.

# Description of Dataset

This is a CSV (Comma-Separated Values) file with 12 features and 64374 entries. The size of the data small but enough to validate the objective of this paper. The data can server the purpose of training an ML (Machine Learning) model for identifying customer patterns and customer churn. This is by identifying the customers that are most likely to churn from the business. The businesses can then choose to retain the customers depending on the churn pro activeness. The author of the dataset is Muhammad Shahid Azeem and is licensed by GPL 2

The features are categorized into categorical and numerical data types. Categorical data refers to the data which cannot be measure and has no numerical values. Categorical data is a grouping like a person can be either “Male” or “Female”. Categorical data can further be categorized into nominal and ordinal. Nominal data does not have numerical values while ordinal data is based on the principle order. In our dataset, there is no ordinal data but rather nominal; gender, Subscription Type and contract type. Numerical data also known as quantitative data is based on numerical values. It can further be categorized into continuous and discrete. The discrete data is finite while continuous data is finite. Continuous data can also be a range. There are no continuous features in our dataset but discrete. The discrete features are CustormerID, Age, Tenure, Usage Frequency, Support Calls, Payment Delay, Total Spend, Last Interaction, and Churn.

Churn is the target variable which is in binary, that is, 1 (yes) and 0 (No). The other numerical features are integers. For categorical data, the data is grouped into several groups. Gender is into two, either Male or Female. “Subscription Type” is in 3 groups; basic, standard and premium. If we were to rank them, “basic” would be in the lowest level, “standard” for middle class and “premium” advanced level. The “Contract Length” is grouped into 3: “Annual” meaning that the contract has a length of 1 year, “Monthly” meaning the contract is for only one month and “Quarterly” is in interval of 3 months.

# Prepossessing

Data prepossessing a crucial part in any machine learning project. In most cases, raw data has missing values, data inconsistencies, typo errors, missing some important columns, and dummy variables. All these problems are sorted during data prepossessing. Data prepossessing methods include, data cleaning, transformation, reduction, integration, normalization, and handling imbalanced data.

Data cleaning involves handling missing values and detecting and removal of outliers. Handling missing values involves replacing “NA” or “Null” values with one of statistical methods or completely deleting it from the data. In statistical methods, we can use mean, median, mode, standard deviation, min, and max to replace data.

Mean refers to average of all values. Thus, if there are n values, then average will be (n0 + n1 ...n4) / n. The mode refers to the most repetitive number in a data value. Mode can be to replace both numerical and categorical values. Median is the middle number in a collection. Min is the smallest number in a collection while max is the largest number in a collection. Any of these methods can be used at a time to replace the missing values. There are no missing values in our dataset, thus none of the above method was used.

Outliers refers to the data points that are furthest from the other data points. Outliers affect mean by either pulling it up or down depending on where they are located. This can in turn affect the accuracy of the trained model. Thus, it is important to remove outliers. Outliers can be detected by use of data visualization techniques such as use of scattered plot, swarm plots and violin plots. These points can be removed by filtering data frame to contain only the data that we require to work with. Let’s take a look at the following list of values:

score = [1, 2, 4, 6, 3, 5, 7, 8, 10, 73, 100]

From above collection, we notice that, the last 2 numbers are very big and they appear as an outcast. These two numbers are the outliers and to remove them, we can filter only the score values that are below 15. Another way of removing outliers is use of scallers such as Max Scaler or Standard Scaler. A scaler is built in a way that it can only accept truncate values that are below 25th percentile and above 75th percentile. The remaining data is believed to be compact and without the outliers. Also, scalers scale data to fit between 0 and 1 which is very efficient when working with machine learning models. In this project, I applied StandardScaler on X features.

Data transformation include scaling, feature engineering and categorical data encoding. Feature Engineering involves adding new features that are sentimental to analysis and model training at large. Feature engineering can be done by combining two or more existing features. Categorical data encoding involves converting categorical data into the numerical values. This can be done by use of dummy algorithm, one-hot encoding or label encoding. We can also replace categorical variables by writing an algorithm that replaces values with the values of our choosing. This can be in the form of a dictionary then applying a map function on a specified column to replace the values. An example of categorical encoding is as below:

Feature “gender” has two unique values “Male” and “Female”. When categorical encoding is applied, then the values will be 0 for “Male” and 1 for “Female”. The reason why we convert categorical data with numerical values is because most of the machine learning algorithms require numerical data. In this project, there is no instance of feature engineering or encoding.

Data reduction is another prepossessing technique that involve use of reducing or sampling data. The two techniques used here are dimensionality reduction and sampling. Dimensionality reduction involves reducing features while preserving the most important ones. An example of dimensionality reduction algorithm is PCA (Principal Component Analysis). On the other hand, sampling involves working with a sample of data instead of the whole data. This is important when dealing with large dataset. There is no instance of dimensionality reduction or sampling in this project.

Lastly, dealing with typo on both feature and values. There is instance where during data collections, instead of typing “Female”, “Fmale” was typed. In such a cases, we find out that instead of having two categories in gender column we have three, that is, “Male”, “Female”, “Fmale”. It is thus logical when we replace “Fmale” with correct spelling. Column names may have spaces in between, for example, “Customer ID” instead of “CustomerID”. There are cases where columns do not have the correct data types. For example, there could be an object data type in column “price” instead of float. This should also be handled to make sure everything is as it should be.

When all of the above has been done the dataset, then the data is clean and ready to be fitted on a model.

# DM Area and DM Algorithm

Apriori Algorithm is a data mining algorithm that is relies majorly on the principle of association. It is applied on the datasets that are contain transaction. Our dataset contains transaction especially in the column “Total Spent” which makes this algorithm for this project. The Apriori Algorithm works by checking strength of the connection between two objects. This algorithm since it generalizes an item reducing search space making it very suitable for large datasets. In this example, we are working with real-world dataset, and scalability should be one of the factors to consider. Apriori is also suitable for providing interpretable results which is very crucial factor when looking for the best strategy to associate the market with. Therefore, Apriori is by far the best data mining algorithm suitable for this project.

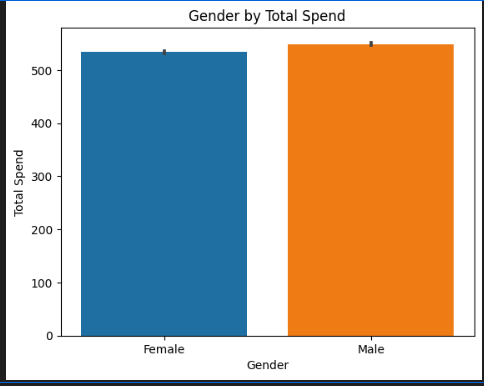
# Exploration of the Effect of k in k-means clustering

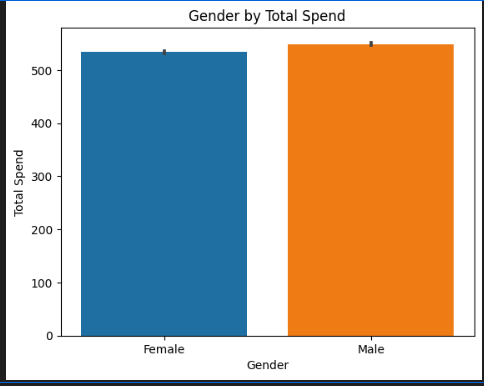
K value affects points association and generalization differently. This is because on different places where the centroids are located. If we use 2 clusters, 2 centroids are formed and the features are pulled towards theses centroids depending on how close they are to each other and similarity in features. The same case applies to the 3 clusters, 4 clusters and so on. There will be a clear line of separation separating the clusters which shows how connected these clusters are. Data mining comes into play when checking the strength of these features to one another. If a feature is weak on one cluster, it is strong on another and thus, it is associated with this another cluster.

# Results and Analysis

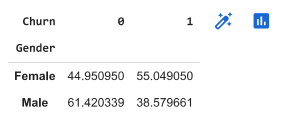
After conducting data analysis on the dataset, the following were the observations:

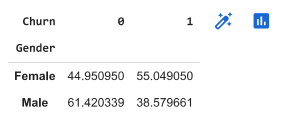
1. There is no much difference between the amount spend by both genders. This is as seen in the bar plot below:

Figure 1: Gender by Total Spend

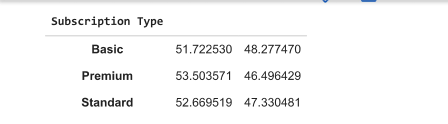
Figure 2: Gender by Total Spend

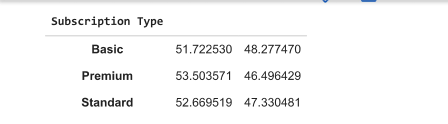
1. Female have more customer churn than male as shown in the cross-table below:

Figure 3: Cross-table for Gender by Churn

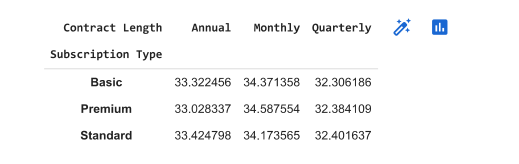
Figure 4: Cross-table for Gender by Churn

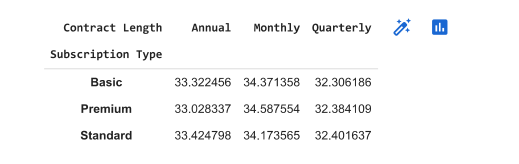
1. Premium subscribers are more, followed by basic then standard.

Figure 5: Customer Subscription Type

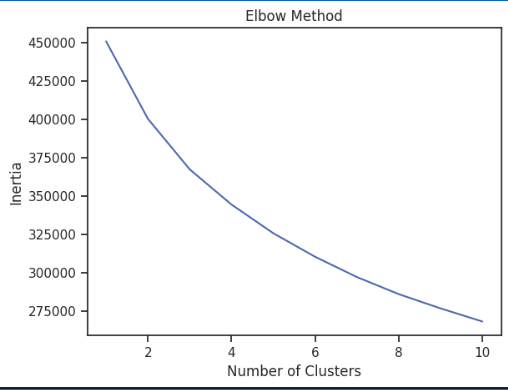
Figure 6: Customer Subscription Type

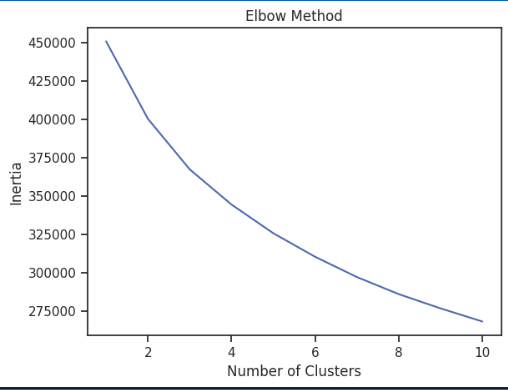
1. Although there seems to be very slight difference in terms of contract length across all subscribers, most subscribers have a monthly contract.

Figure 7: Contract Length by Subscription Type Cross-table

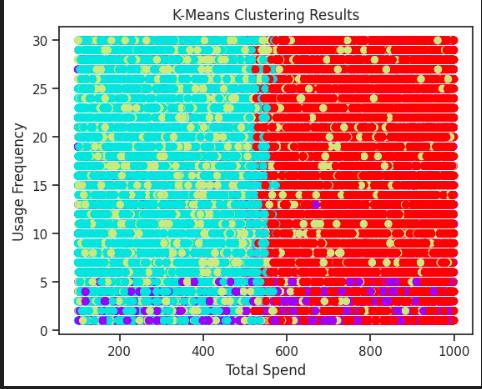
Figure 8: Contract Length by Subscription Type Cross-table

1. When we use elbow method to determine the optimal k value, we get 4 clusters. This can be shown as follows.

Figure 9: K-Elbow

Figure 10: K-Elbow

This can be further be visualized below when the features are classified which shows clearly that there are 4 clusters.

Figure 11: K-means clustering results

# Conclusion

K-means has proved to be an effective unsupervised classification algorithm in machine learning. To ensure no stone is left un-turned and acquire the best results, features are subjected to various clusters and the best is chosen from the results presented. The clusters can be either manually selected or by use of k-elbow graph and the clusters on the elbow are chosen. For this dataset, the features can be put into 4 clusters.

We have also seen that Apriori Algorithm is one of the best when working data mining project due to its scalability, efficiency and the fact that it can present interpretable results. All these factors are necessary for predicting market trends and helping in implementation of the best strategy that favors the current market.

# References

Customer Churn Dataset. (n.d.). Www.kaggle.com. <https://www.kaggle.com/datasets/muhammadshahidazeem/customer-churn-dataset?select=customer_churn_dataset-training-master.csv>

*K means Clustering - Introduction - GeeksforGeeks*. (2019, May 30). GeeksforGeeks. https://www.geeksforgeeks.org/k-means-clustering-introduction/

Pérez-Ortega, J., Almanza-Ortega, N. N., Vega-Villalobos, A., Pazos-Rangel, R., Zavala-Díaz, C., & Martínez-Rebollar, A. (2019). The K-means algorithm evolution. *Introduction to Data Science and Machine Learning*.