Understanding why customers fail to continue using the services of a certain business helps decision makers to implement policies that encourage customers to continue using the services of a company. Given the importance of customer attrition or churn, numerous past researches have been conducted, notably, due to the advanced techniques of collecting huge data and application of machine and deep learning algorithms, it now possible to predict more precisely the reasons behind customer churn. One of the methods used to predict customer churn is the k means clustering algorithm. This brief literature review will use therefore review literature on k means clustering, customer churn, and how k means clustering algorithm can be used to predict the customer churn.

Research by Xiahou & Harada (2022) applied different methods on e-commerce data to develop efficient customer retention strategies. The researchers used support vector machine (SVM), logistic regression and k-means clustering algorithm. Notably k-means clustering was used as a precursor step to applying SVM and logistics regression. The reason the authors choose to use k means clustering as a precursor is because k-means is unsupervised while the latter methods are supervised. Unsupervised learning helps researchers to group observations into clusters that have the same characteristics, hence it does not have labelled data for the training dataset. The obtained model from the training dataset is then used on the test dataset to determine how accurate the model used is. Research by Matuszelański & Kopczewska (2022), revealed that accuracy of a model can be determined by the area under the ROC curve, the bugger the area the more accurate the model is, additionally, the researchers found that a confusion matrix also surfaces as an effective tool of measuring the accuracy of the fitted unsupervised model. Findings from Matuszelański & Kopczewska (2022) fitted models on customer churn revealed that factors such as the location of the customer, the first purchase reviews, the amount paid for the first order, shipping and category of the product purchased, and customer’s demographic environment were the main factors causing customer churn in a retail e-commerce store in Brazil. Findings from the two outlined studies reveal that there is no one machine learning model that fits all, hence the process of fitting a machine learning model should be interactive.

In the case of k-means clustering, there exists several rules that guide the clustering procedure, according to Sinaga & Yang (2020) one of the rules is determining the number of clusters, that the data should be segmented into. For instance, the number of segments can be from 2 to or even 10, depending on the goals and nature of the dataset. Notably, there are other advanced methods of determining the optimal number of clusters, one of the most commonly used method is the elbow method. The elbow method involves iterating k = 1 to k = n where the within-cluster sum of squares (WCSS) is calculated for each k Cui (2020). With this method, it is possible to consider very many different numbers of k’s in a single instance. It is called the elbow method because that when a graph of WCSS is plotted against k, the elbow point is the optimal number of clusters to be used when implementing the k-means algorithm. Notably, there exist other methods of choosing the number of k clusters such as the Silhouette Method, Davies-Bouldin Index, and Hierarchical Clustering Dendrogram (Schubert, 2023).

The main advantages of using k-means clustering algorithm is the ease of in implementing the algorithm, can be scaled to huge datasets, it is easy to visualize the algorithm hence making data interpretation easier, does not consume huge computation power. As outlined earlier, k-means can be used as a precursor to other supervised machine learning methods such as SVM and logistic regressions. Notably, there exists several disadvantages of suing k-means clustering, the first disadvantage is choosing k clusters manually.

References

Cui, M. (2020). Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance*, *1*(1), 5-8.

Matuszelański, K., & Kopczewska, K. (2022). Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*(1), 165-198.

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE access*, *8*, 80716-80727.

Xiahou, X., & Harada, Y. (2022). B2C E-commerce customer churn prediction based on K-means and SVM. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*(2), 458-475.

Schubert, E. (2023). Stop using the elbow criterion for k-means and how to choose the number of clusters instead. ACM SIGKDD Explorations Newsletter, 25(1), 36-42.