Customer Segmentation in E-Commerce

[1]Vani Ashok, [2]Rahul R Kamath, [3]Adithya RK, [4]Supreeth Singh,[5]Ajay Bhati

[1]Assistant Professor, [2]B.E Student, [3]B.E Student, [4]B.E Student, [5]B.E Student
[1]Dept. of Computer Science and Engineering,
[1]JSS Science and Technology University, Mysuru, India

Abstract : Customer segmentation is the process of dividing customers into groups based on common characteristics so that companies can target each group efficiently. With the increase in businesses coming up every day, it has become significantly important for the old businesses to apply marketing strategies to stay in the market as the competition is increasing. This project is an attempt to classify the customers into different categories using unsupervised learning and train supervised learning classifiers using this data and decide the best classifier based on the values of accuracy. In this project many machine learning libraries present in Python are used to perform different operations. KNN, Random Forest, Gradient Boost are the supervised learning algorithms used and an accuracy of 68.29, 75.50, 75.77 was achieved respectively during testing.

Keywords: **Segmentation, KNN, Random Forest, Gradient Boost**

1. INTRODUCTION

With the growth of the market, it has become mandatory for the old businesses to apply marketing strategies to stay in the market due to the increase in competition. The number of consumers is increasing significantly every day and it has become challenging for the businesses to cater to the requirements of every and **e**ach customer. Hence Customer segmentation is used which is the process of dividing customers into groups with respect to the common characteristics by which the companies can target each group efficiently. It allows companies to see what actually the purchasers are buying which can prompt the businesses to better serve their customers leading to customer satisfaction, it also allows the businesses to seek out who their target customers are and improvise their marketing tactics to get more revenues from them.

2. LITERATURE SURVEY

Li, Zeying [1] have proposed a method in which a retail supermarket was taken as research object, and data mining methods was used to retail enterprise customer segments, and then association rules obtained using Apriori algorithm were used to different groups of customers and get rules about customer characteristics to make customer characteristic analysis efficiently. Finally, the author gave some references to the supermarket's marketing and management work, which helped in understanding it in detail. Data mining was used efficiently to deal with the large number of historical and current data, from the database to find some potential, useful and valuable information for the retail stores which help us target customers.

Wang, Zhenyu, Yi Zuo, Tieshan Li, CL Philip Chen, and Katsutoshi Yada [2] have analyzed customer segmentation based on broad learning system which provides an alternative view of learning in deep structure. Firstly, in addition to customer purchasing behavior, RFID (Radio Frequency Identification) data was also included, which can accurately represent the consumers' in-store behavior. Secondly, this paper used Broad Learning System (BLS) to analyze the consumer segmentation. BLS is one of the finest machine learning techniques, and quite efficient and effective for classification tasks. Thirdly, the customer behavior data used in this paper was collected from a real-world supermarket in Japan. Customer segmentation was considered as a multi-label classification problem based on both of POS data and RFID data.

Kansal, Tushar, Suraj Bahuguna, Vishal Singh, and Tanupriya Choudhury [3] performed customer segmentation using K-means clustering. A python program was developed and the program was trained by applying standard scaler onto a dataset having two features of 200 training sample taken from local retail shop. Both the features are the average of the amount of shopping by customers and average of the customer's visit into the shop annually. By applying clustering, 5 segments of cluster were formed labelled as Careless, Careful, Standard, Target and Sensible customers. However, the authors got two new clusters on applying mean shift clustering labelled as High buyers and frequent visitors and High buyers and occasional visitors.

Bhade Kalyani, Vedanti Gulalkari, Nidhi Harwani and Sudhir N Dhage [4] have proposed a systematic approach for targeting customers and providing maximum profit to the organizations. An important initial step was to analyze the data of sales acquired from the purchase history and determine the parameters that have the maximum correlation. Based on respective clusters, proper resources can be assigned towards profitable customers using machine learning algorithms. K-Means clustering was used for customer segmentation and Singular Value Decomposition was used for providing appropriate recommendations to the customers. This paper also deals with the drawbacks of the recommender system like sparsity, cold start problem etc and how they can be overcome.

3. RESEARCH METHODOLOGY

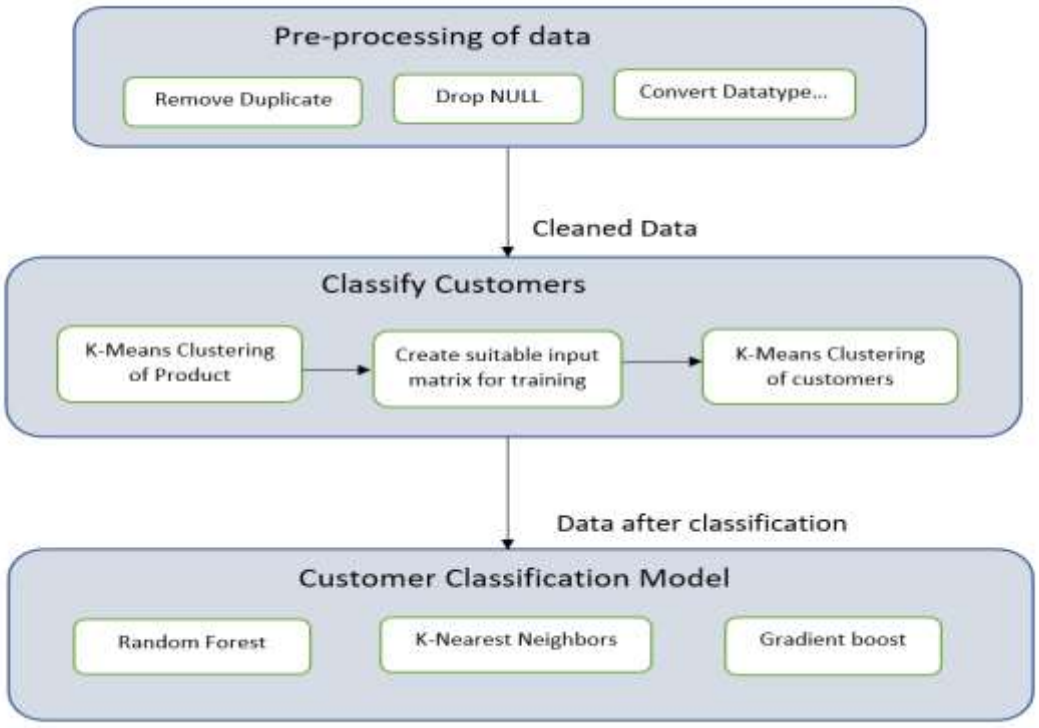


Fig 1: System architecture for customer segmentation.

Fig 1 shows the system architecture for customer segmentation. The data set obtained is passed to the pre-processing stage where duplicates and null rows are dropped. A dataset may contain cancelled orders which are removed as well. The datatype of columns is converted to the required datatype for pre-processing. The refined data set is then used to classify the customers into different categories by un-supervised learning through K-Mean algorithm. The resultant dataset or matrix which contains customer and its categories is used to train and test supervised training models such as Random Forest, K-Nearest Neighbors, and Gradient Boost.

3.1 Data Preprocessing:

The dataset was obtained from a machine learning repository in University of California [5]**.** Initially it was found that the dataset has 5,41,909 rows and 8 columns (Invoice No, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID, Country). After removing all the NULL values, the data frame was of 4,06,829 rows and 8 columns. It was observed that almost 25% of customer ID entries were NULL. It was then observed that there were 5,225 duplicate rows in the data frame. It was then observed that there were 37 different countries in the dataset and plotted a choropleth map for the same showing the orders per country. It was observed that 16.47% of the orders were cancelled. Some entries had negative values for quantity so it was assumed that when an entry is cancelled in the dataset, an entry with same Customer ID, description is added but with a negative value of quantity so it was decided to locate the entries. The dataset had discount entries, which got discarded later. Even then the dataset had some negative quantity entries which proved our assumption to be wrong. It was because the dataset has values from 2010 but the remaining values are those which were made before 2010 but cancelled after 2010.

3.2 Classifying customers using unsupervised learning:

A basket was created and a basket price list for every transaction and a pie chart for different range of values was plotted. It was seen that ~65% of purchases give prizes in excess of £ 200.In the next step analysis of product categories was performed. A function was defined that takes as input the data frame and analyses the content of the Description column. Firstly, it extracts root word corresponding to a word using stemmer function and the values associated with each root word and represented it using a dictionary (Key, Value). It then counts the number of times each root appears in the data frame. When several words are listed for the same root, it was considered that the keyword associated with this root is the shortest name (E.g. if we have runner and running as values associated with root word run, runner will be selected as it has smaller length). A subplot showing the number of occurrences of every word was plotted. The keywords were used to create group of products and defined the matrix X with product along y axis and description along x axis. Entry in the matrix will be 1 if the description of the corresponding product contains the word. The product was then grouped into classes using K-means and silhouette score was calculated which indicates how well the clusters have been formed. +1 indicates perfectly formed clusters, -1 indicates overlapping clusters. It was realized that 5 clusters will be perfect in our case as the increase in silhouette score after 5 wasn't significant.

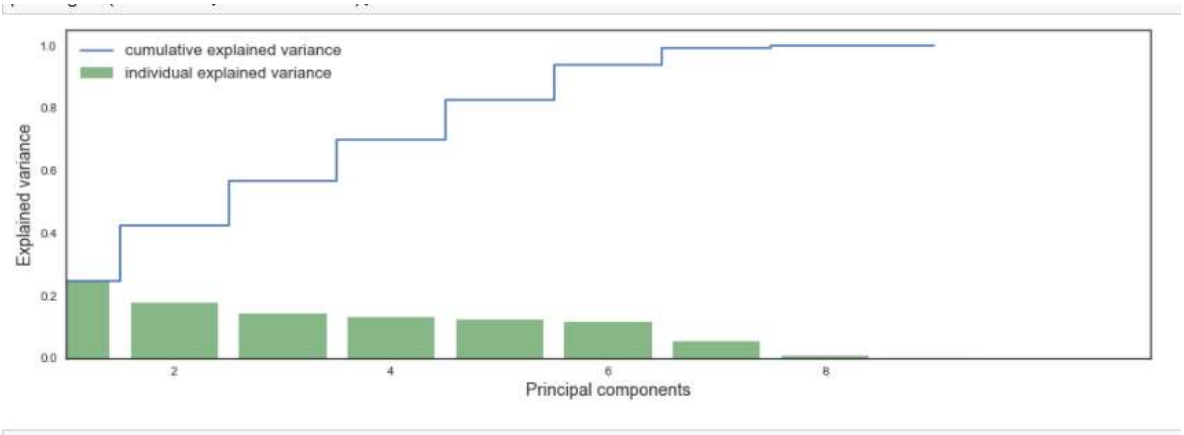
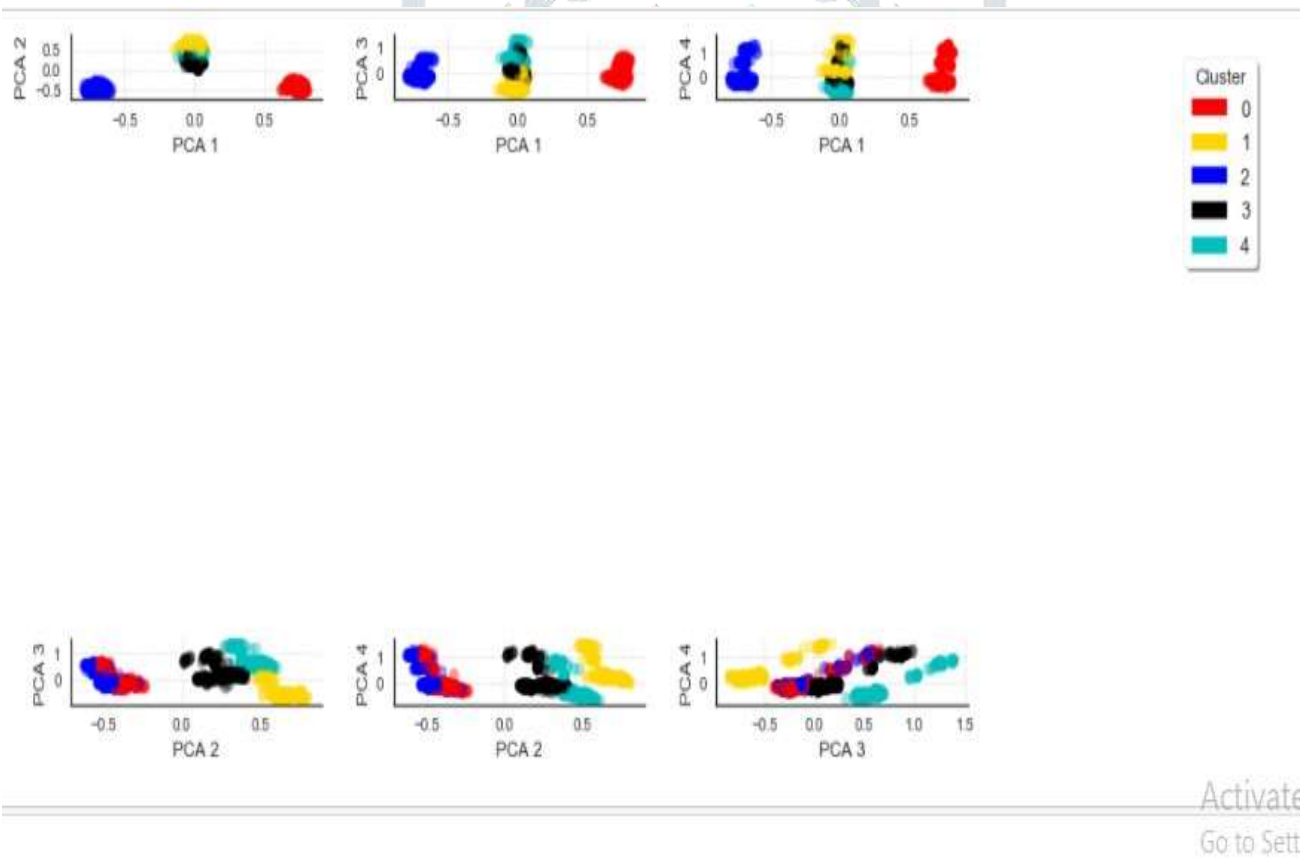


Fig 2: Projecting PCA to convert the feature data from a higher dimensional space to a lower dimensional space.

In the Fig 2, a graph was plotted with principal components along the X axis and explained variance along the y axis. Explained variance which is the variance shown by individual components, increased with the increase in the number of principal components. Hence, it was observed that, as the number of principal components increases, loss decreases. Mean, sum, count, min, max for each customer and percentage of the customer being allocated that each category was calculated.

Standard Scalar was used to standardize features by removing the mean and scaling to unit variance. Hence, a matrix with standardized values was generated. In practice, before creating the clusters of customers, it is interesting to define a base of smaller dimension allowing to describe the matrix. In this case, this base was used in order to create a representation of the different clusters and thus verify the quality of the separation of the different groups. Therefore, PCA was performed beforehand.



From the Fig 3, it was concluded that many clusters are well separated while some are overlapping. At this point, clusters of clients were created from the standardized matrix that was defined earlier and using the K-means algorithm a silhouette score of 0.218 was obtained.

3.3 Steps in supervised learning:

A model was built by feeding the obtained data to various supervised learning classifiers and accuracy for every classifier was obtained and results were compared. Generally, determination of whether a given model is optimal is done by looking at its F1: Score, precision, recall, and accuracy (for classification), or it's coefficient of determination (R2) and error (for regression). However, real world data is often distributed somewhat non uniform, meaning that the fitted model is likely to perform better on some sections of the data than on others. Cross Validation Scores visualizer enables us to visually explore these variations in performance using different cross validation strategies. Training score indicates how well the model is fitted for the training data. The 3 supervised classifiers used are Random Forest, K nearest neighbors, Gradient Boost. A random forest is a estimator that fits a number of decision tree classifiers on different sub-samples of the dataset and uses averaging to improve the calculated accuracy and control over-fitting. The sub-sample size is controlled using the max_samples parameter if bootstrap=True (default), else the whole dataset is used to build each tree [6]. In k-NN classification, the output has a class membership. An object is classified by a majority vote of its neighbors, with the value being assigned to the class most common among its k nearest neighbor [7]. Gradient Boost (GB) builds a model in a forward stage-wise fashion; it allows the user to optimize the values of arbitrary differentiable loss functions. In each stage regression trees are fit on the negative gradient of the binomial or multinomial deviance loss function. Binary classification is a special case in which only a single regression tree is induced [8]. The 3 different classifiers Random forest, K-NN, and Gradient Boost resulted in accuracy of 75.50, 68.29, 75.77, respectively for testing. A confusion matrix is a table which gives information about the performance of a classification model on validation data for which the true values are known. The accuracy of the results seems to be as expected. Nevertheless, when the different classes were defined, there was an imbalance in size between the classes obtained. In particular, one class contains around 40% of the clients. Hence, we need a confusion matrix.

4. RESULTS

Fig 4 shows the learning curve for Random Forest. From the learning curve, it can be concluded that both the training score and cross validation score increases initially and reaches a constant value. An accuracy of 90.30, 75.50 was achieved for training and testing of Random Forest respectively.

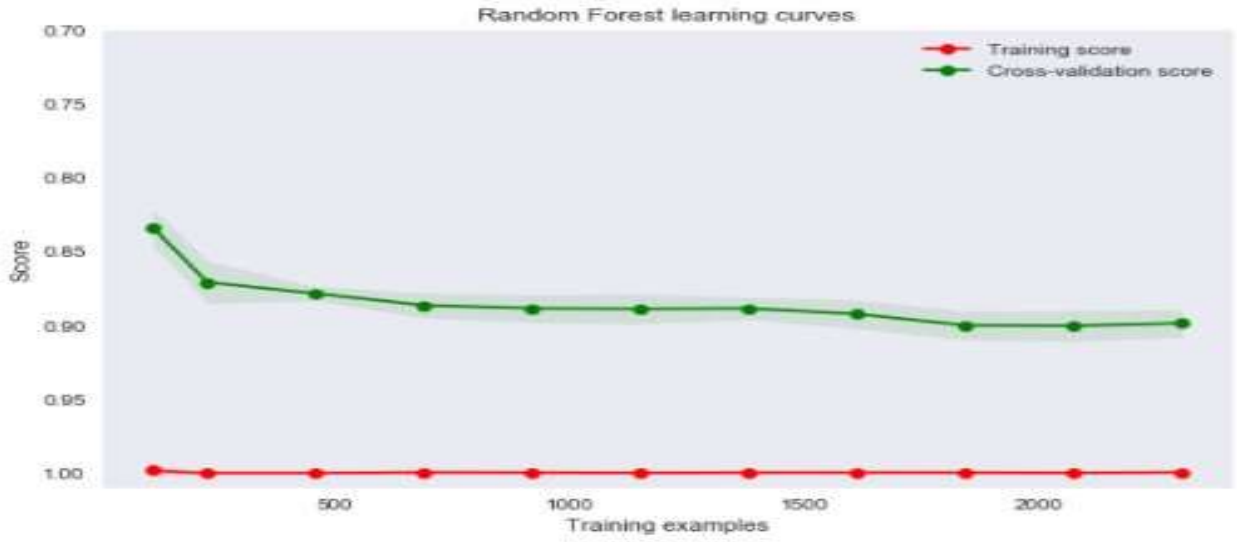


Fig 4: Learning curve for Random Forest.

Fig 5 shows the learning curve for Random Forest. From the learning curve, it can be concluded that the training score decreases with increase in samples but cross validation score increases with increase in samples. An accuracy of 90.17, 75.77 was achieved for training and testing of Gradient Boost respectively.

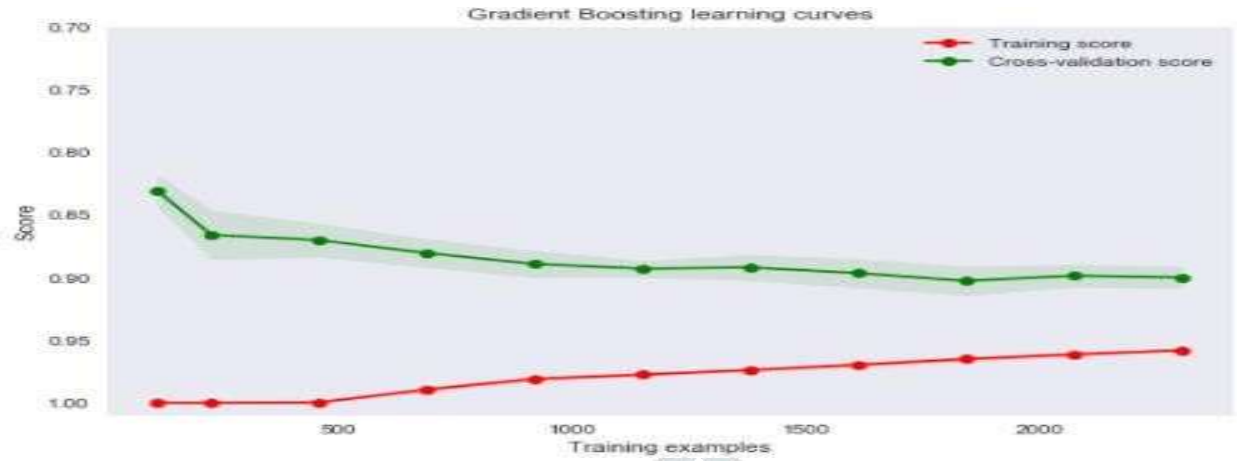


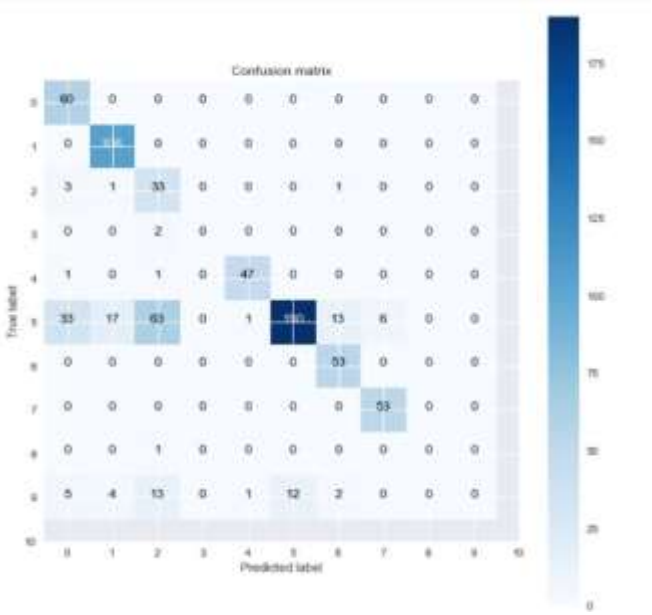
Fig 5: Learning curve for Gradient Boosting

Table 1: Training and testing accuracy

CLASSFIER	TRAINING ACCURACY	TESTING ACCURACY
K-Nearest Neighbors (k=5)	81.44	68.29
Random Forest	90.30	75.50
Gradient boost	90.17	75.77

Table 1 shows the training and testing accuracy for the 3 different classifiers. A good accuracy of 75.50,75.77 was achieved for Random Forest and Gradient Boost respectively whereas for KNN, accuracy of 68.29 was achieved which is comparatively low.

Fig 6 is the confusion matrix for the 11 customer categories. From the confusion matrix, the cases where the predicted class was same as the true class and the cases where predicted class was different from the true class can be known. The entries along the diagonal represent the number of correct predictions and the other entries represent the number of incorrect predictions. Hence most of the entries are correctly predicted.



5. CONCLUSION

The proposed work aims to classify online e-commerce customer into various categories based on their characteristics like spending amount, type of product they buy, and how frequently purchase happens etc. Initially the products were classified into 5 categories using K-means and through this result along with other characteristics like amount spent, frequency etc. online e-commerce customers were classified into 11 categories using un-supervised method i.e. K-means Clustering. To measure the accuracy, silhouette scoring was used and a score of 0.02 was obtained. PCA was used to obtain insights about the clusters. Later different classifiers were trained using the data obtained so far. The 3 different classifiers Random forest, K-NN, and Gradient Boost resulted in accuracy of 75.50, 68.29, 75.77, respectively. The values of accuracy for Random Forest and Gradient Boost are almost same. Further from the learning curve of those two, it can be verified that both of them neither overfit nor underfit. Hence both the classifiers are equally good for the given dataset.

REFERENCES

[1] Li, Zeying. "Research on customer segmentation in retailing based on clustering model." In 2011 International Conference on Computer Science and Service System (CSSS), pp. 3437-3440. IEEE, 2011

[2] Wang, Zhenyu, Yi Zuo, Tieshan Li, CL Philip Chen, and Katsutoshi Yada. "Analysis of Customer Segmentation Based on Broad Learning System." In 2019 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), pp. 75-80. IEEE, 2019.

[3] Kansal, Tushar, Suraj Bahuguna, Vishal Singh, and Tanupriya Choudhury. "Customer segmentation using K-means clustering." In 2018 international conference on computational techniques, electronics and mechanical systems (CTEMS), pp. 135-139. IEEE, 2018.

[4] Bhade Kalyani, Vedanti Gulalkari, Nidhi Harwani and Sudhir N Dhage "A Systematic approach to customer segmentation and buyer targeting for profit maximization" In 2018 9th International Conference on Communication and Networking Technologies(ICCCNT), pp. 1-6 IEEE,2018.

[5] https://archive.ics.uci.edu/ml/datasets/Online+Retail

[6] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[7] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

[8] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html