RESEARCH ARTICLE OPEN ACCESS

Implementation of Customer Purchase Prediction using Machine Learning Techniques

Ms. N. Zahira Jahan M.C.A., M.Phil.,*M. Gokul** *
Associate Professor/MCA, Department of MCA, Nandha Engineering College, (Autonomous), Erode, Tamilnadu, India. Email: zahirajahan1977@gmail.com
** Final MCA, Department of MCA, Nandha Engineering College, (Autonomous), Erode, Tamilnadu, India. Email: gokulmurugan06@gmail.com

****

Abstract:

Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. One such task is to help make predictions on online shopping data. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform delivery time reduction with higher accuracy by using different machine learning techniques. The main motivation of doing this project is to present a delivery time reduction model for the sales company. Further, this research work is aimed towards identifying the best classification algorithm for identifying the reduction time in delivery. Online retailers still struggle with the disadvantage of delivery times compared to traditional brick and mortar stores. With the emergence of big data analytics, it has become possible to extract meaningful knowledge from the volume of data that online retailers collect on their website. Nevertheless, limited research exists that investigates how this data can be used to optimize delivery times for customers. Different forecasting methods in combination with k-means clustering are applied to test if, and how early, it is possible to predict online purchases. Results indicate that customer purchases are, to a certain extent, predictable, but anticipatory shipping comes at a high cost due to wrongly sent products. The proposed prediction model can easily be implemented and used to predict purchases, which can also be leveraged for other areas of application besides anticipatory shipping Model development is based on Means clustering and categorization using Support Vector Machine algorithms along KNearest Neighbour also. It is found that proposed machine learning algorithm performs better when compared to other algorithms for delivery time reduction. The project is designed using R Language 3.4.4 with R Studio.

Keywords — **Machine Learning K-Means,KNN Algorithms , SVM Algorithms.**

****

I. INTRODUCTION

Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. One such task is to help make predictions on online shopping data. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform delivery time reduction with higher accuracy by using different machine learning techniques. The main motivation of doing this project is to present a delivery time reduction model for the sales company. Further, this research work is aimed towards identifying the best classification algorithm for identifying the reduction time in delivery. Online retailers still struggle with the disadvantage of delivery times compared to traditional brick and mortar stores. With the emergence of big data analytics, it has become possible to extract meaningful knowledge from the volume of data that online retailers collect on their website. Nevertheless, limited research exists that investigates how this data can be used to optimize delivery times for customers.

II. LITERATURE STUDY

The online shopping has become the regular a part of today's world. The factors related to this shopping trend are Convenience, Better Product Selection and Useful Delivery Mode. There exist many pros and cons of online shopping and through their study this and future trends of online shopping are often analysed. Data mining is additionally an efficient field to investigate consumer behavior in online shopping. The buying patterns of customer are somehow interlinked. Association rule mining helps identifying such patterns and helping the business decision making process. The behavior of shoppers shopping over the web has been analysed through five factors. These factors are time, privacy, trust, convenience and merchandise variety. The research conducted for this purpose was within the kind of questionnaire and the results were analysed statistically. Trust was claimed as such a person's trait that affects their buying habits. The shopping behavior of individuals in Pakistan gets affected by their psychological and emotional attitudes. Privacy is additionally considered because the most prominent consider this online shopping trend within the way that folks may sometimes not feel secure sharing their personal information over the net. On the other hand, fascinating prices of assorted stuff may additionally attract individual attention and aid to urge online shopping. So, trust over the source is that the issue that affects people buying behaviours. To identify consumer behavior a hunt has been conducted on 7-Eleven shop. The factors considered during this regard are cultural, social, personal and psychological factor. After applying multivariate analysis and hypothesis testing the coefficient of determination (R2) springs that describes the influence of all the factors that affects consumer inclination to shop for the products. of these factors also are independently discussed and analysed statistically. The online reviews help the shoppers establishing an opinion regarding online shopping. These reviews vary in both quality and quantity. they will have both positive and negative kind of effect on customers and moreover as on business. The data during this regard is collected via questionnaire and also the results are compiled after various complicated statistical methods. These results interpret that reviews do contribute to decision making in online shopping. There are various techniques of knowledge mining for the identification of frequent item sets. because the data retrieved after processing is incredibly large and requires some efficient technique to discover some useful pattern. The paper discusses those techniques which will aid within the formation of any such pattern. Association rule has been considered united of the fundamental data mining tools. There exist many algorithms like KNN Algorithm algorithm, AIS, SETM,KNN Algorithm hybrid, FP-growth for pattern discovery. except the pros and cons of

those algorithms, any of those may be used together with association rule mining for data analysis .

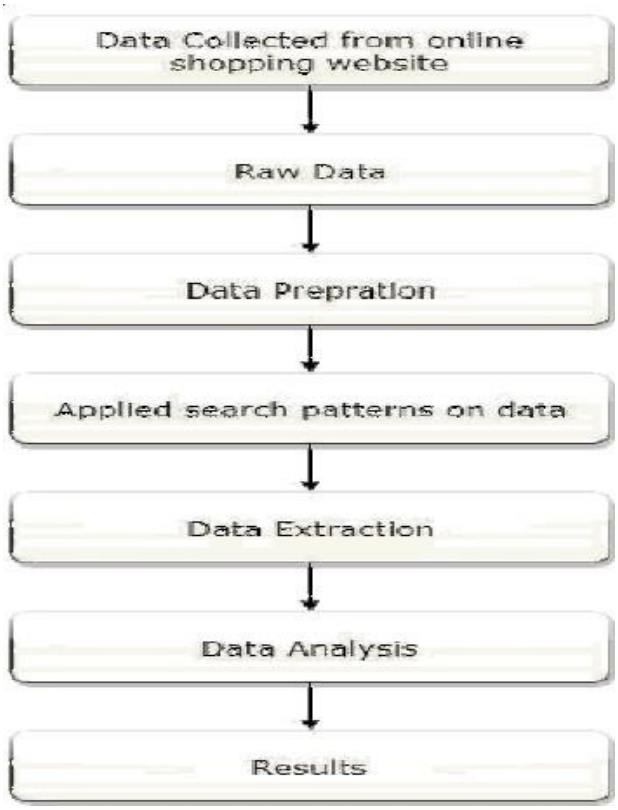

Fig. 1. Flow Chart of Methodology

In this project, to exemplify our research work we used an online shopping home decoration website's database as a sample database. the location is about multiple items like Duvet covers, Bedspread, Curtains, Cushions, pillows and loo and kitchen accessories. the location contains thousands of records that can provides a little insight about customer purchasing habits in home decoration. the explanation for selecting this database was that, it's a working site and has such a large amount of numbers of customers visiting a day that gives a large dataset to analyse customer behaviour as Fig. 1 presents a technique of our work.

III. PROPOSED WORK

All the existing system approaches are carried out in proposed system. In addition, KNN algorithm is used to predict the model as it helps better in various ways. It is found to be suitable especially if the data set is having a greater number of records is contains outlier data. A wide variety of shopping records can be taken for classification purpose and predicting a new model at the same time increasing the efficiency.

IV. METHODOLOGY

1. K-Means Clustering

The K-NN algorithm is that the simplest among all other machine learning algorithms'-NN is one amongst the most known supervised learning algorithm in pattern classification. In KNN, the output may be a class membership. for every information, the algorithm finds the K nearest observation, so classifies the information point to the most. K nearest neighbour's are found near the query point element. Calculate distance between query point and every element of matrix. The distance which is a smaller amount that's considered to be nearest neighbour. To calculate the space, Euclidean Distance method is employed. The advantage of KNN is that the price of learning process is zero, no assumptions about the characteristics of the concepts to learn must be done and therefore the complex concepts can be learned by local approximation using simple procedures. it's a non-parametric and lazy learning algorithm. In KNN makes prediction using the training dataset directly. In KNN k is that the number of nearest neighbour and it's the deciding core factor. Generally most of the k values are in odd numbers because it contains classes that are in even numbers. The given below graph contains two classes namely class A and class B. For this k value is one. Consider C is that the point, for which label has to predict. Find the closest point to the C then classify points by majority vote of its k neighbour's. Each object votes for his or her class and therefore the class with the bulk votes is taken into account because the prediction. for locating the foremost nearer point calculating the gap between the points using distance measures like Euclidean distance, Manhattan distance, Hamming distance and

Minkowski distance. KNN has some three basic steps first step is to calculate the space second step is to finding the neighbour elements and final step is to vote for labels. The K-Means clustering is a type of unsupervised learning used when in unlabelled data (i.e., data without defined categories or groups).K-Means algorithm is considered as one of the most popular, reliable and effective algorithm. It is usually used with a least squared distance error to identify clusters depending on the specified number of clusters. The similarity distance of two elements is calculated by Euclidean distance, Manhattan distance, Mahalanobis distance, Hamming. The Euclidean distance is calculated using the formula

$$d(x, y) = \sqrt{\sum_{i-1}^{n} (x_i - y_i)^2} \quad (1)$$

In order to tackle with the categorical variables, the K-Modes algorithm is usually used. Again, K-Modes cannot compute for continuous variables. So, the K-Prototype algorithm is introduced for both categorical and continuous variables. K-Prototype inherits the ideas of K-Means, it applies Euclidean distance to numeric attributes of the closeness between two objects.

K-Means clustering of bank customer data are used to analyse and group the customers. The customer data set is taken with Customer, Total Orders, Total Amount, etc columns of which Total Orders and Total Amount are taken as X and Y Axis for KMeans clustering. By default 5 is given for K, but we can give any number to cluster the data.

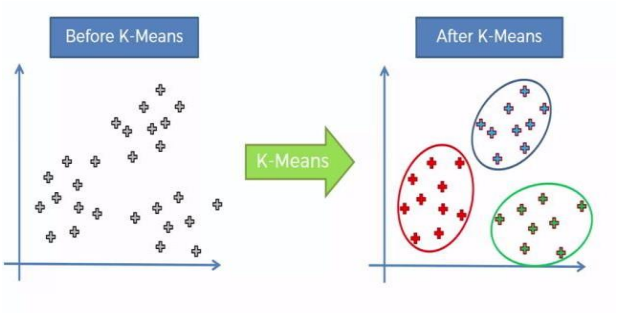

Fig 2. K-Means Clustering

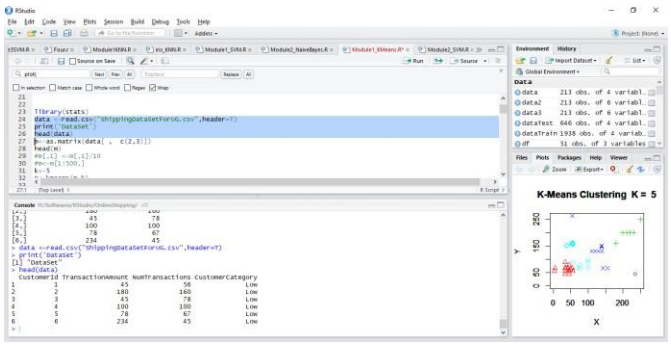

Fig 2.1 Dataset records

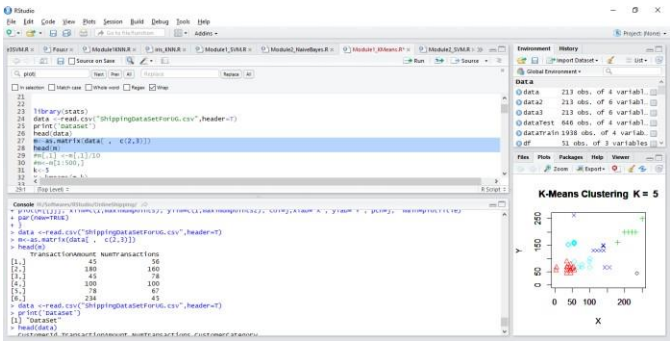

Fig 2.2 K-Means Columns

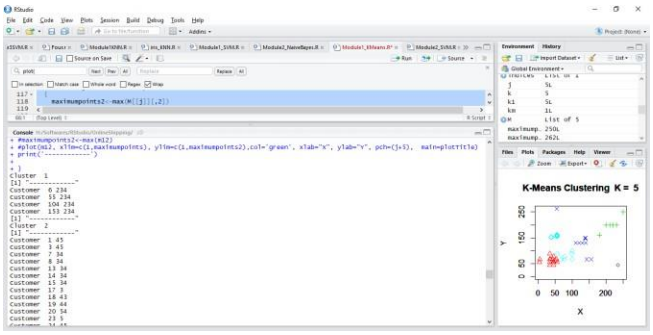

Fig 2.3. Customers Clustered

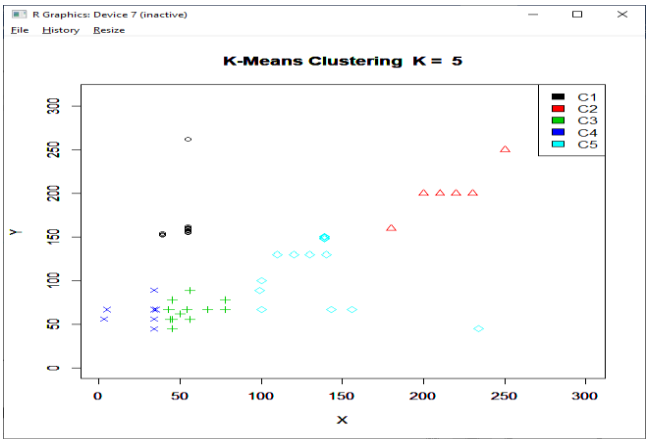

Fig 3. K-Means Output of Shipping Data

2. Support Vector Machine (SVM)

¨Support Vector Machine (SVM) ¨ might be supervised machine learning algorithmic rule which may be used for every classification or regression challenges. However, it's principally utilized in classification issues. During this algorithmic program, it tend to plot every knowledge item as a degree in n-dimensional with worth of each feature being the price of a specific coordinate. Then, it tend to perform classification by finding the hyper - plane that differentiate the 2 categories.

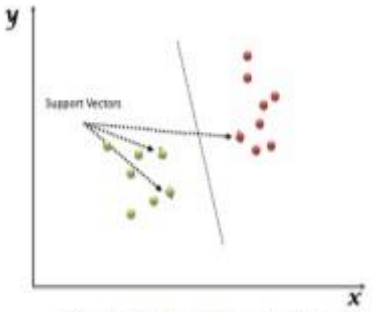

Fig4. Model Graph of SVM

Support Vectors are merely the co-ordinates of individual observation. Support Vector Machine could be a frontier that best separates the 2 categories (hyper-plane/line). Application of SVM includes text and electronic text categorization, as their application will considerably cut back the requirement for labeled coaching instances in each the standard inductive and transductive settings. The advantage of SVM is that it works rather well with clear margin of separation and it effective in high dimensional spaces. The SVM algorithm is to search out the hyperplane in an N dimensional space. Hyperplanes are the choice boundaries which classifies the info values. The hyperplane dimension fully depends upon the number of features.

$$f(x) = B(0) + sum(ai * (x,xi))$$

This is an equation that involves calculating the inner products of latest input vectors(x) with all support vectors in training data. The coefficients BO and ai (for each input) must be estimated from the training data by the educational algorithm.

3. K Nearest Neighbour Classification (KNN)

KNN ensures a machine learning technique on the basis of statistical learning theory. KNN is applied with 'K' parameter, training data and testing data. 75% training records and 25% testing records are given to predict the model. The customers are classified as Low value or high value customer. In addition, each customer record is checked against all remains records and near matching records of count 'K' is found out and classified accordingly. KNN Algorithm is based on feature similarity: How closely out-of-sample features resemble our training set determines how we classify a given data point.

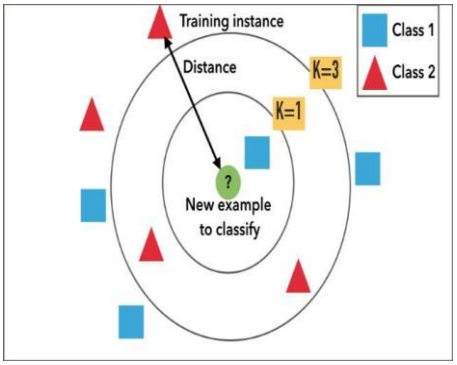

Fig 5. KNN Value

KNN can be used for classification — the output is a class membership (predicts a class - a discrete value). An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours. It can also be used for regression - output is the value for the object (predicts continuous values). This value is the average (or median) of the values of its k nearest neighbours.

V. DATAFLOWFLOW DIAGRAM

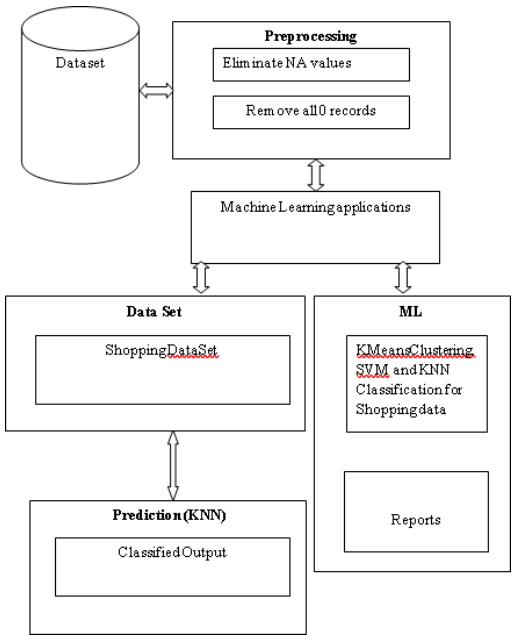

Fig 6. DataFlow Diagram

VI. RESULTS

After applying multiple queries, the results obtained are shown within the tables above. These tables are quite self explanatory in the sense that they supply the acceptable answer to the 4 questions raised by us, for the aim of research. The KNN Algorithm algorithm further support the answers in this regard providing a tip to marketing strategy within the analysis of customer behavior in an exceedingly way that the products that are closely related together, in terms of use or offered in a very deal together are more of an opportunity to be bought together.

VII. CONCLUSION & FUTURE WORK

As it is discussed above target of this project is to analyse behavior of such people who are visiting the online shopping sites and spending their time there, checking different stuff. We have taken the database of a running website related to the online shopping. Number of products and categories related to the product is present in it. We mined the data from database, ultimate tool used and queries are applied to extract the data. In future Advance KNN Algorithm can be applied to observe the customer behavior analysis through support and confidence of product mining association rule.

VIII. REFERENCES

[1] J. H. a. P. Gendall, "Understanding and predicting human behaviour," Wellington , 2008.

[2] S. M. Prof. Sanjeev Kumar, "Status and Scope of Online Shopping: An Interactive Analysis through Literature Review," International Journal of Advance Research in Computer Science and Management Studies , vol. 2, no. 12, 2014.

[3] R. Abhijit Raorane, "Data mining techniques: a source for consumer behavior analysis," 2011.

[4] D. S. Sven F. Crone, “Predicting customer online shopping adoption an evaluation of data mining and market modelling approaches”.

[5] J. C. O. J. Paul Peter, “Consumer behavior and marketing strategy, McGraw-Hill”, 2005.

[6] I. M. W. K. B. Dr Rizwana Bashir, "Effects of online shopping trends on consumer-buying behavior: an emparical study of pakistan," Journal of Management and Research, vol. 2, no. 2, 20\15.