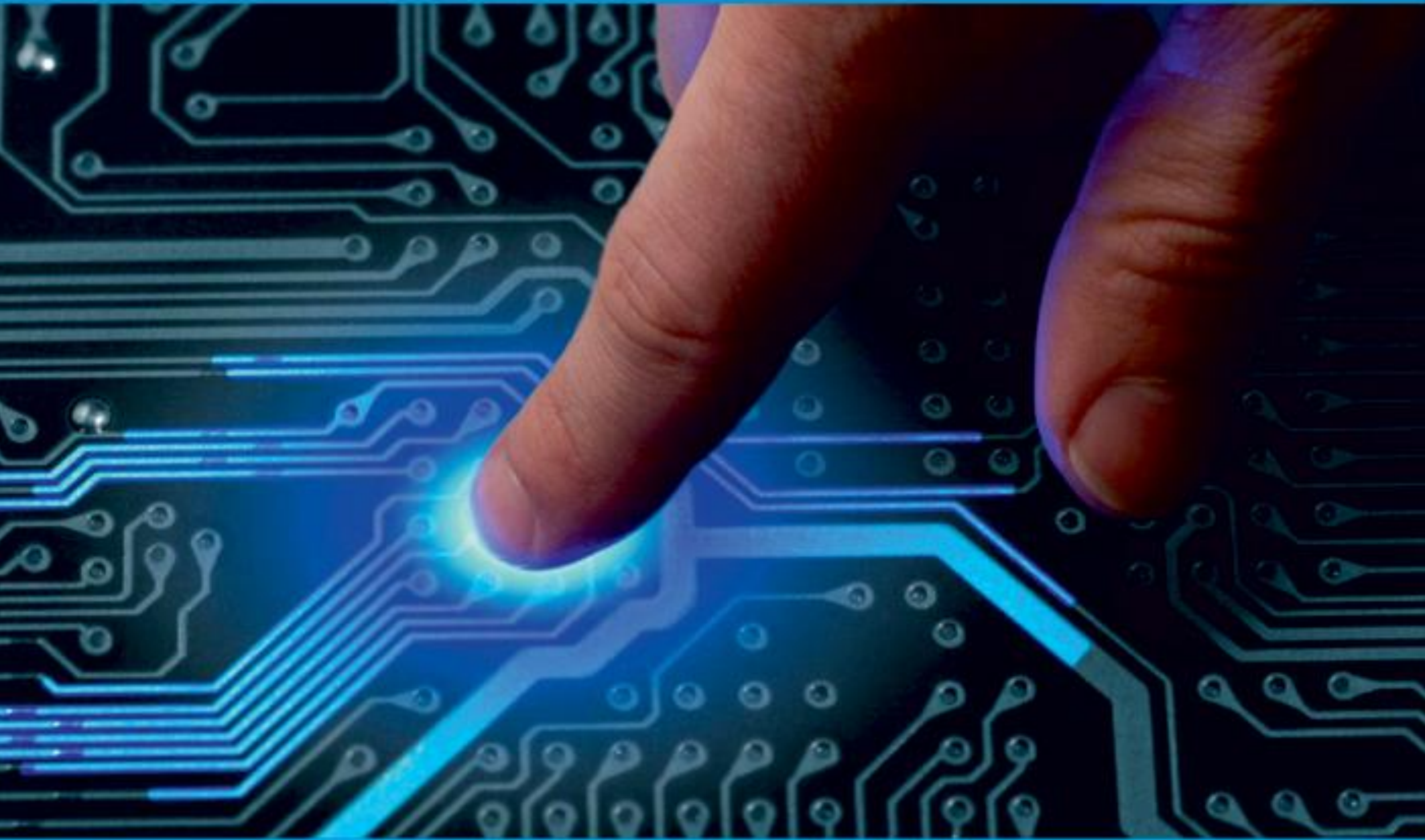
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 6, June 2024

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.379

Product Distribution Analysis using Data Mining Techniques

Sagar Haldar

Department of Computer science and Information Technology, Chhatrapati Shivaji Maharaj University, Panvel, Navi Mumbai, Maharashtra, India

ABSTRACT: Over the last few decades there is an exponential increase in raw data that is needed to log, process and analyse in order to reduce the wastage of storage space as well as loss of hidden information. Here the concept of Knowledge Discovery in Databases(KDD) and data mining is used in order to retrieve useful information from these raw data. Retails and services are the major sector of development. A large amount of data is generated on daily basis. When this data is processed and analysed, we can forecast various aspects that can be used to manage resource wisely to maximize profit while at the same time minimize the cost. This paper is aimed at providing a review of implementation of real world application of data mining techniques in Retail data in order to discover hidden information to forecast the distribution of products to increase profitability.

KEYWORDS: Data mining, Retails and services, Regression, Classification, Clustering

I. INTRODUCTION

The process of extracting interesting and useful information from raw chunks of data is known as data mining, which is otherwise manually impossible, as it consumes a lot of time and human efforts. The concept of data mining was first introduced in 1990 by database community. In 1989, Gregory Piatetsky-Shapiro phrased the term “Knowledge Discovery in Databases” in Knowledge Discovery in Databases workshop. Over time scientist has related the term KDD to datamining, this is due to the fact that knowledge is the product of data mining. As of today datamining and Knowledge discovery are used interchangeably. Today data mining is used in various fields such as retails and services, medicines and healthcare, education etc. Retail industry is one of the major sector which is growing big everyday. India is one of the fastest growing retail market in the world. It is worth $1.2 trillion in FY 2023. With this, ever growing market industries has invested a lot to research the various aspects such as consumer shopping behaviour, customer loyalty and satisfaction, price perception, services-scapes, store atmosphere, distribution of products etc. Retailers plans to insure success or maximize profit by learning about these factors that affects their sales. When there is sudden spike in sales and if the retailer is caught off-guard there might not be enough stocks in particular store. With insights to the sales of products in a particular store, the retailer can make better allocation of products to these stores which will maximize the profit in time with minimum cost.

The main focus here is to analyse the retail data and to find out how the product sales is going on in particular retail stores, which is further used to forecast the distribution of products into the retail stores that is scattered in various geographical location.

II. RELATED WORK

In 2015, Harsoor & Patil worked on forecasting sales of Walmart store using big data applications: Hadoop, MapReduce and Hive so that resources are managed efficiently. In 2017 Manpreet & Mahmood used the same concept to gain new insights into the consumer behavior to comprehend Walmart’s marketing efforts and their data-driven strategies through visual representation of the analysed data.

III. BACKGROUND WORK

3.1 Problem Studied

Retailers often tries to find out how to distribute the products into retail shop in order to maximize the profit. They also have to make sure that the products are sold out by the end of the season or month without leaving any excess items. There are many factors that can affect the sales of particular product from a retail store such as season, area, customer’s preferences, selling price etc. The retailers have to take these into account before distributing the products. This problem is solved here by using data mining techniques.

3.2 Tools and techniques applied

The process includes the collection of retail sales datasets stored in CSV format. The collection of data is organized into named columns. Python programming language used. Pandas library is used to process and transform these data. Matplotlib and Seaborn library is used to visualize the datasets. Scikit-learn library is used for machine learning and data modelling.

IV. DATA MINING TECHNIQUES USED

The various Data mining algorithms used in this project are regression, classification, clustering.

4.1 Regression

Regression in Machine learning is a type of supervised method. In Machine Learning Regression gives the relationship between independent variables(features) and dependent variables (outcome or target variable). In Regression, algorithms are trained to understand the relationship between the independent and dependent variable. The model is then used to predict the outcome of new input data. In regression algorithm is used to predict continuous outcomes. There are mainly two types of regression namely: Linear regression and logistics regression. The output of regression analysis gives the relationship strength of variables, which is known by correlation coefficient. The correlation coefficient gives value between -1 and 1, which tells how strongly or weakly the variables are related.

4.2 Classification

Classification is a type of supervised method in Machine learning. Classification method is used to predict the correct label of the given input. In classification the algorithms classify or forecasts the object in predefined classes such as true or false, male or female, present or absent etc. In classification algorithm, a discrete output function(y) is mapped to input variable(x), which is given by equation

$$Y=f(X), \text{ where } Y = \text{categorical output}$$

There are mainly two types of classifiers:

1.Binary classifier: If the classification problem has only two possible outcomes, then it is called as Binary Classifier. Example: Yes or No, Male or Female, Spam or Not spam.

2.Multi-class Classifier: If a classification problem has more than two outcomes, then it is called as multi class classifier.

Example: Classification of animals, Classification of music.

Both Classification and Regression algorithms uses labelled data. But the key difference is that the regression works where the data has continuous or numerical value, whereas classification predicts discrete labels or categories.

The figure below shows how the target variable defines the selection of supervised algorithms

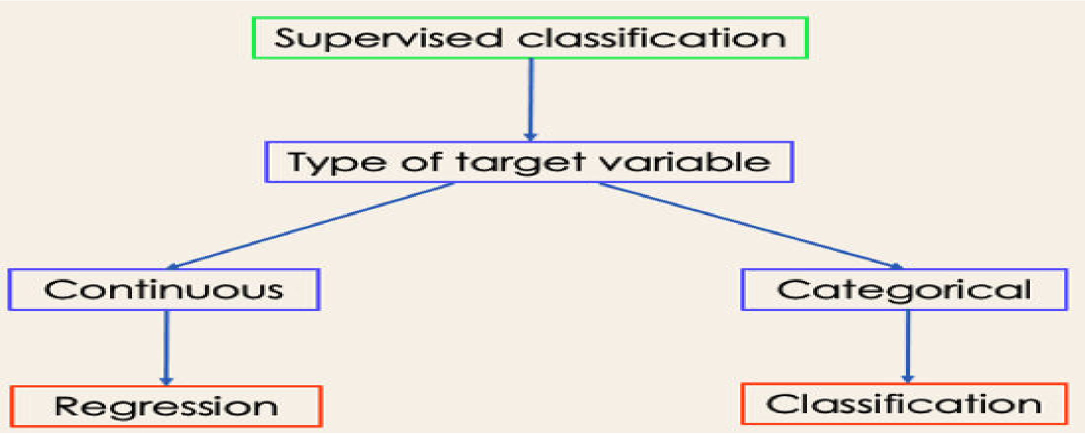


Figure 1-Model Selection

4.3 Clustering

Clustering is a type of unsupervised method in Machine learning. It uses unlabelled data. In clustering the algorithms identifies clusters and make groups of similar object based on the similarity of features of that object. The algorithm is then used to predict that in which cluster the input object belongs. Unlike supervised learning we don't have target variable in clustering. Clustering aims at forming groups of homogeneous data point from heterogeneous dataset. It evaluates the similarity of data points based on metric like Euclidian distance, Cosine similarity, Manhattan distance, etc. and then groups the points with highest similarity scores together.

There are mainly two types of clustering:
1.Hard Clustering: In this type of clustering, each data points belongs to a particular cluster either completely or not. For example, if there are 4 objects and we have to cluster them into two separate clusters, then each objects either belongs to cluster 1 or cluster 2.

2.Soft Clustering: In this clustering, instead of assigning each object to a particular cluster, a probability or likelihood of that object is calculated. This is often applied to the objects where the features of an object do not fulfil required criteria. For example, if all the features of an object are not given then the algorithm will calculate probability of the object belonging to a particular cluster based on the available features.

There are many clustering algorithms like Hierarchical clustering, Gaussian Mixture Model(GMM), Spectral clustering, Mean Shift clustering, etc. The most common clustering algorithm used is k-means, which is also used in this project.

V. IMPLEMENTATION

5.1 Data

The data is collected using OLTP (Online Transaction Processing) API and is stored in CSV format. The data is stored in named columns. This raw and unprocessed data include categorical data, numerical data and text data in raw form which is processed later.

5.2 Importing Libraries/modules

Essential python libraries such as Pandas for data processing, scikit-learn for Machine Learning, seaborn and matplotlib for data visualization is imported.

5.3 Data pre-processing

The data collected includes null values, spaces, special characters etc., which is then processed as follows :

- The first step is to remove null values. These null values are filled with 0. Then it is replaced by the mean of their respective columns using Pandas libraries.
- Special characters and space in text data is removed using regular expressions.
- At last data is converted from text to numerical values as the machine learning model accepts numerical value only.

5.4 Applying ML Algorithms

After cleaning or pre-processing the data the next step is to implement ML algorithms. Various ML algorithms are used such as Logistic regression for finding relationship between variables, which is described as correlation coefficient, that is visualized later in heat map, Decision tree and Random forest classifiers are used for classification purpose, and k-means clustering algorithm is used for making clusters of products that describes how the product is to be distributed in various stores in various locations.

5.5 Code Implementation

The entire implementation is done in Python language with several packages and modules.

Vectorization: Using scikit-learn's 'labelEncoder' along with fit_transform method to convert the cleaned data into numerical vectors.

Data Splitting: Data is splitted for model training and testing in ratio of 7:3 respectively using scikit-learn's 'train_test_split' method.

ML Training: The data is trained using scikit-learn's 'LogisticRegression', 'RandomForestClassifier' and

'DecisionTreeClassifiers' for correlation coefficient and classification respectively. For Clustering scikit-learn's 'KMeans' is used.

Data Visualization: For visualization of output result matplotlib and seaborn is used. Heat map is generated in order to show correlation coefficient, which is further used to analyse data using Principle Component Analysis(PCA) and is used reduce dimension or remove unnecessary columns. Scatter graph is generated using matplotlib's 'pyplot' for showing product cluster distribution.

VI. RESULT AND DISCUSSION

The following are the result of the project:
1)Heat map generated is shown below:

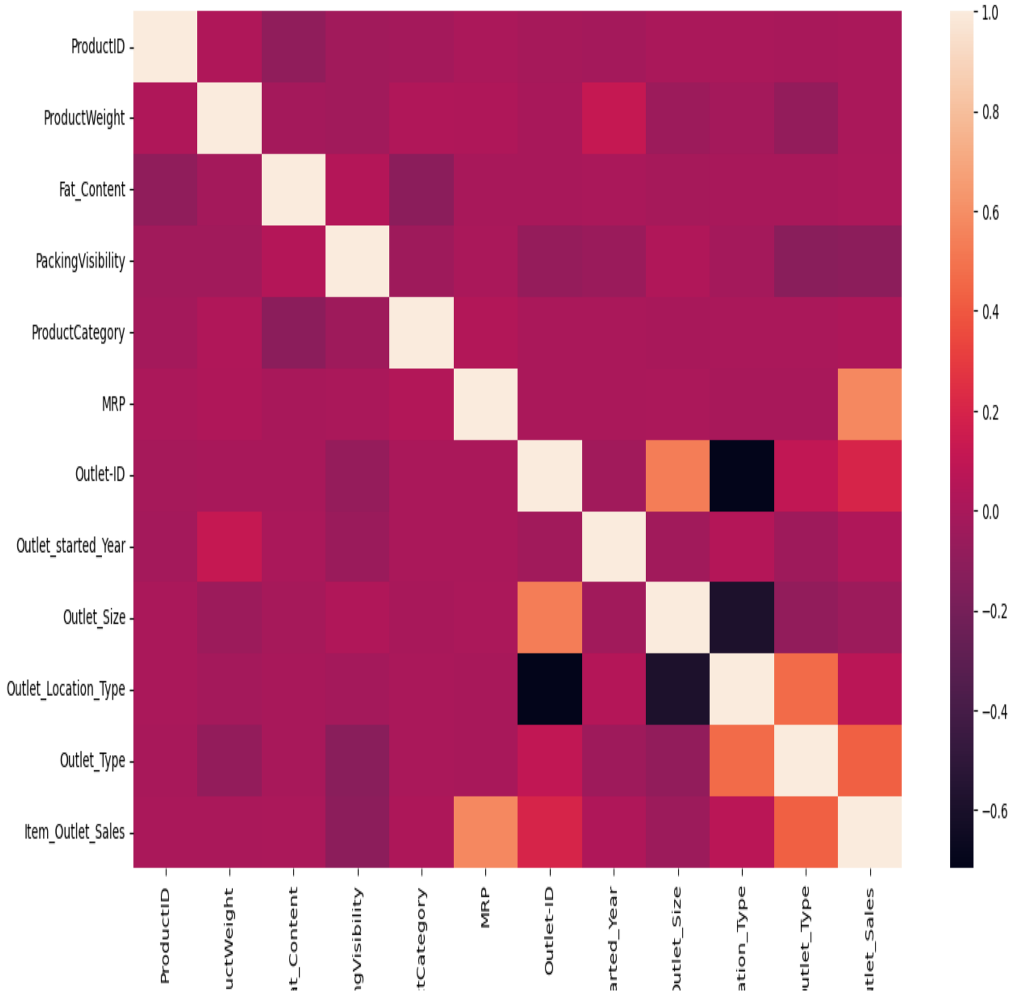


Figure 2-Heat Map of Correlation Coefficient

2)The Heat map shows the correlation coefficient of all the columns. It shows the relationship between variables of particular columns, with 1 as strongest relationship and -1 as the weakest relationship. As it is visible some columns have weakest relationship, these columns can be removed manually or the dimensions can be removed by PCA, as this data is irrelevant.

3)After dimension reduction the data left is used for cluster formation that describes how the product is to distributed in the stores of a particular geographical location. The scatter graph generated is shown below.

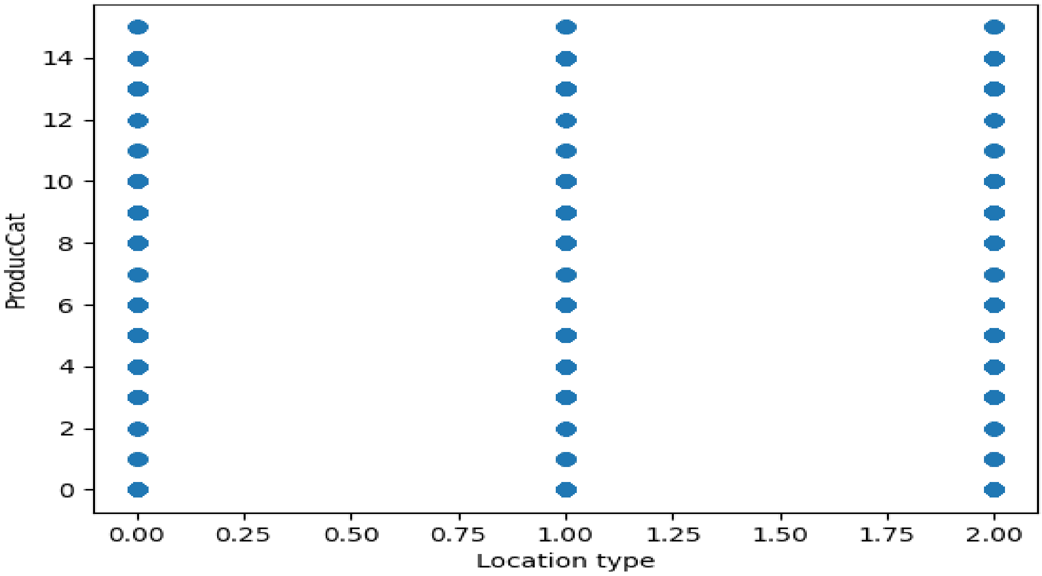
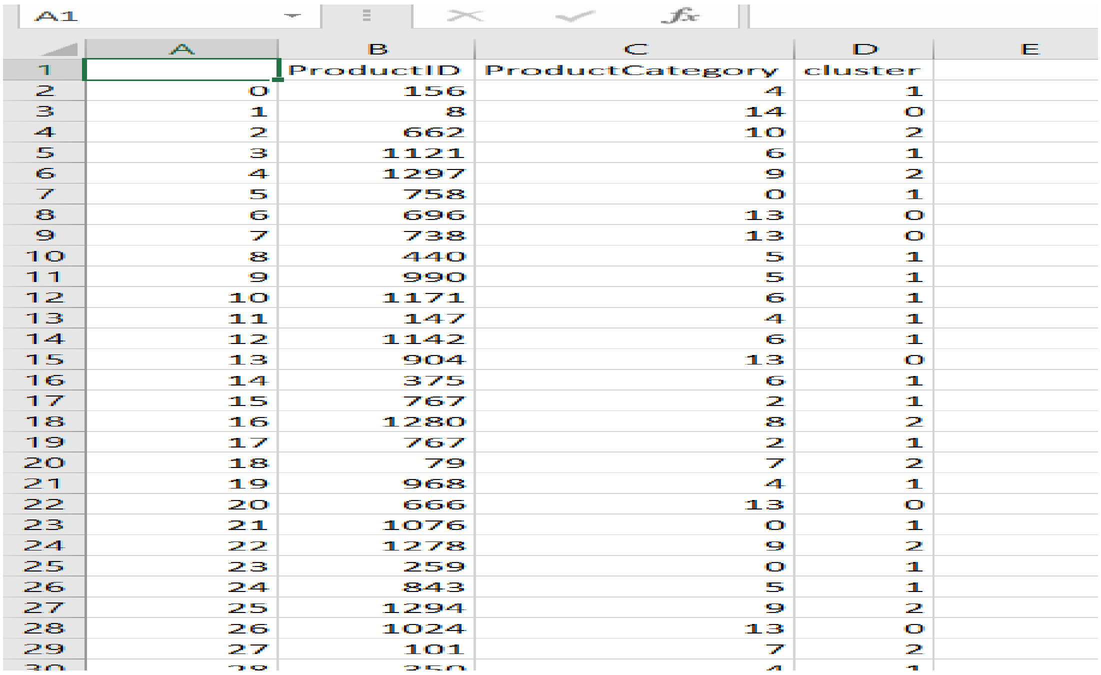


Figure 3-Scatter plot of product distribution

As shown the product can be distributed in three areas which is represented as 0,1,2.
4)Furthermore the product that is clustered is saved in CSV file, which is easily interpreted and can be used for distribution of product in the given three stores. The data saved is shown below:



	A	B	C	D	E
1		ProductID	ProductCategory	cluster	
2	0	156	4	1	
3	1	8	14	0	
4	2	662	10	2	
5	3	1121	6	1	
6	4	1297	9	2	
7	5	758	0	1	
8	6	696	13	0	
9	7	738	13	0	
10	8	440	5	1	
11	9	990	5	1	
12	10	1171	6	1	
13	11	147	4	1	
14	12	1142	6	1	
15	13	904	13	0	
16	14	375	6	1	
17	15	767	2	1	
18	16	1280	8	2	
19	17	767	2	1	
20	18	79	7	2	
21	19	968	4	1	
22	20	666	13	0	
23	21	1076	0	1	
24	22	1278	9	2	
25	23	259	0	1	
26	24	843	5	1	
27	25	1294	9	2	
28	26	1024	13	0	
29	27	101	7	2	
30	28	250	4	1	

Figure 4-CSV file of product distribution

VII. CONCLUSION

In conclusion, the ever growing sector of retail market needs to managed efficiently. And as the years passes retailers will need to understand the customer needs and manage the resources wisely. It can be done by gaining useful insights of the customer behaviour by mining the combination of products bought by them. This can give the rough estimation of the budgets required and sales made by different stores for a period of time, further allowing data outliers to be defined and their characteristics to be listed. With the evolving technology this process is reaching out to new levels, new Data mining techniques is taking over the traditional method of managing and analysing data. These technologies are constantly used to understand complex datasets in a matter of time with beautiful visual representations. The visual representations gives an easy way to understand and interpret the data effectively. Through observing these datasets, clearer ideas on the sales on particular retail shop is interpreted which is very helpful to the market on its own. Additionally, seasonality trend and randomness and future forecasts will help to analyse sale drops which the companies can avoid by using a more focused and efficient tactics to minimize the sale drop and maximize the profit with minimum cost and loss. It can further be built upon using higher level research.

REFERENCES

[1] A. S. Harsoor and A. Patil, “Forecast of sales of walmart store using Big Data application,” International Journal of Research in Engineering and Technology, vol. 4, p. 6, June 2015.
[2] J. Dean and S. Ghemawat, MapReduce: simplified data processing on large clusters. Association for Computing Machinery, 2008.
[3] Manpreet Singh and Mahmood Rashid: Walmart’s Sales Data Analysis - A Big Data Analytics perceptive,2017.
[4] W. J. Frawley, G. Piatetsky-Shapiro and C. Matheus, "Knowledge Discovery in Databases: An Overview," AI Magazine, vol. 13, no. 3, pp. 57-70, 1992.
[5] L. A. Kurgan and P. Musilek, "A survey of Knowledge Discovery and Data Mining Process," The Knowledge Engineering Review, vol. 21, no. 1, pp. 1-24, 2006.
[6] F. Weiping and W. Yuming, "The Development of Data Mining," International Journal of Business and Social Science, vol. 4, no. 16, pp. 157-162, 2013.
[7] Ian Davey and Technolegis, "Consumers, Big Data, and Online Tracking in the Retail Industry: A CASE STUDY OF WALMART," 10 August 2014.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details