Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 6 Issue 9, September - 2019

Predictive Analysis For Consumer Demand Products In Fast Moving Consumer Goods (FMCG) Using Hadoop Framework

Murari Thejovathi
Research Scholar
Computer Science and Engineering
Acharya Nagarjuna University
Guntur, Andhra Pradesh, India
theju.scient@gmail.com

Dr. M.V.P. Chandra Sekhara Rao
Professor
Computer Science and Engineering
RVR & JC College of Engineering, Guntur, Andhra Pradesh, India manukondach@gmail.com

***Abstract*—Analytics are everywhere and strongly connected to our everyday lives. Big data analytics are used primarily in various sectors for accurate prediction and analysis of the large data sets. They can find an important information from large data sets. In this paper, we are creating a Hadoop based data pipeline for streaming and processing real time data from different FMCG companies include production, sales, quality, marketing, strategies using open source and google cloud. Based on this data, we can perform predictive analysis for consumer highly demand products in their daily life. FMCG products are used by most people day in and day out. Hadoop big-data framework is used to handle large data sets through distributed storage and processing real time data. K-Nearest neighbour classification algorithm is used to classify the data for performing analysis. Finally, we used Machine Learning module of Spark for pre-processing the data because it Integrates well with the Hadoop ecosystem and data sources. Spark Streaming receives the input data and divides the data into datasets and classes, after that we can process by the Spark engine and will generate final stream of results in batches using GraphX.**

Keywords—Big data analytics, Hadoop big-data framework, FMCG Products, classification algorithm, Machine learning Module spark.

I. INTRODUCTION

Big Data' is a term that defines the huge volume of data that is growing exponentially. Data analytics includes extracting the useful information from the data by building all possible relations among various data. This makes big data to appear even bigger. big data is one of the most discussed topics in business today across industry sectors. As one of the most "hyped" terms in the market today, there is no consensus to define big data. The term often used

synonymously with related concept such as Business Intelligence (BI) and data mining. This also forms the basis for the most used definition of big data, the three V: Volume, Velocity and Variety as shown in Figure 1.

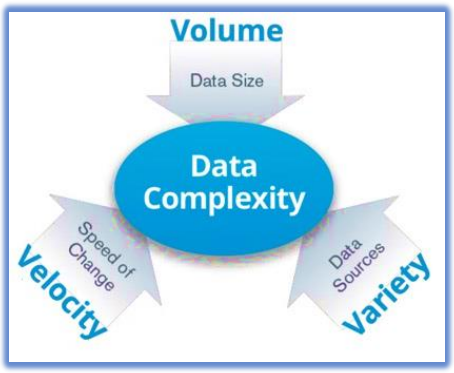


Figure 1: Three V's of Big data

Importance of Big Data Analytics

Big data analytics helps to find solutions for problems like cost reduction, time-saving and lowering the risk in decision making. By combining data analytics and machine learning, the many organisations / companies can gain a lot by:

•Risk management and calculating potential risk causes.

•Determining causes of failure in policies of businesses and eliminating the causes in future.

•Time-to-time offers for the customers based on their purchases.

•Detecting any fraudulent activity using cross-checking of data.

FMCG Fast Moving Consumer Goods

Fast-moving consumer goods are also called consumer packaged goods (CPG). Fast Moving Consumer Goods (FMCGs) are defined as products which are sold quickly at relatively low costs. These are mainly non-durable consumer goods which are required extremely frequently and, in some cases, almost daily by a consumer. [1] *Some of Leading FMCG Companies in India* ITC, Nestle, Colgate-Palmolive, Parle Agro, Britannia Industries, Marico, Procter & Gamble.

The categories of FMCG products shown in Figure 1.1

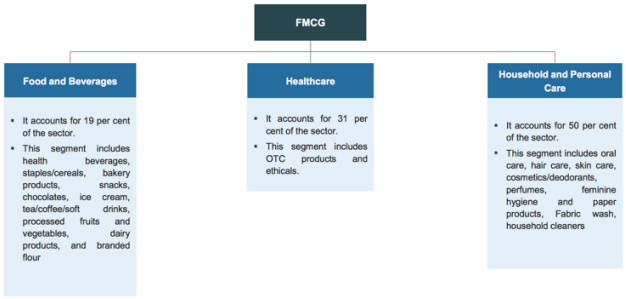


Figure 1.1: Categories of FMCG Products.

The Emerging consumer trends and their future impact on FMCG industry shown in Figure:2

The FMCG sector has grown from US$ 31.6 billion in 2011 to US$ 52.75 billion in 2017-18. The sector is further expected to grow at a Compound Annual Growth Rate (CAGR) of 27.86 per cent to reach US$ 103.7 billion by 2020. This sector will be projected to grow 11-12 per cent in 2019.

Figure:2 Future impact on FMCG Industry

ANALYTICS IN RETAIL & FMCG:

CUSTOMER ANALYTICS

Market Equations in India which helps the organization in Retail and Consumer Goods (FMCG) industry identify and target the right customers through effective customer segmentation, improve acquisition and conversions through personalized offers - product, price, promotion, minimize churn and enhance customer loyalty, retention and life time value through effective cross sell and up sell strategies by identifying the best combination of products and services that align best with the customers' needs.[2]

PREDICTIVE ANALYTICS

Our Analytics Outsourcing services for the Retail and Consumer Goods (FMCG) industry helps organizations to build high quality predictive models and helps to understand the customers propensity to buy/upgrade, develop LTV models, predict churn, forecast demand assisting inventory planning and replenishment and enhance revenue opportunities and profitability. [2] [13].

In this paper we are using the implementation of machine learning techniques on Apache Spark framework for finding results of consumer highly demand products in the FMCG companies, in the form of tabular and graphical representations.

A. Existing System

For each FMCG company has their own data base to store the data. and different FMCG products has their own characters and analysis Here we are collecting different FMCG companies' real time data using google cloud platform for analysing the real time data.

B. Problem statement

In this paper, we are analysing all the different FMCG products real time data through Hadoop framework using Apache spark [3] for finding the consumer demand products and their characteristics, needs to help the FMCG sector growth.

C. Machine Learning in FMCG

The FMCG (Fast Moving Consumer Goods) industry is an ideal target for Predictive Analytics and Machine Learning. [12] [13] There are several unique attributes of the industry that makes the predictions. They are:

- The massive volumes involved
- Access to good quality sales data
- Short shelf life
- Current forecasting techniques are relatively inaccurate
- Current marketing strategies are less than optimal
- Current manufacturing practices are less than ideal
- Current supply chain strategies are less than optimal
- Consumer numbers are very large

II. PROPOSED MODEL

A HDFS based data pipeline has been proposed for streaming and processing the FMCG real time data. In this data, we can perform analysis for finding the consumer demand FMCG products in daily life.

The following is the proposed and created Hadoop based data pipeline model to stream the real-time data of FMCG sector. The detailed model has been shown in Figure:3

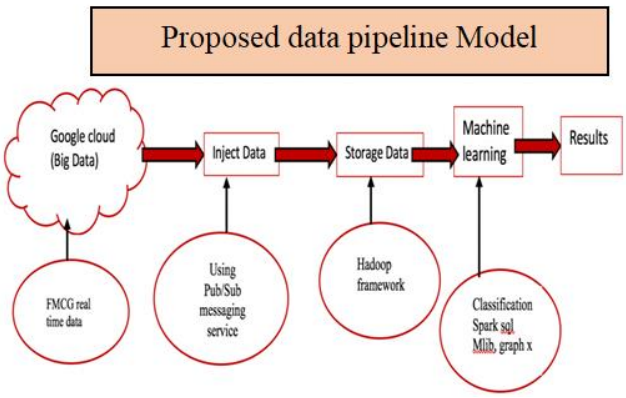


Figure:3 Proposed HDFS based data pipeline

The overall proposed model has been divided into Three phases:

1. Inject the real-time FMCG data from cloud into HDFS framework

2. Implement KNN classification algorithm to split the data into a training dataset.

3. Using Apache spark Pre-process the datasets for finding predictive analysis of consumer demand products from the FMCG industry.

A. Google Cloud Platform and Apache Beam

Google Cloud Platform provides a group of useful tools for big data processing.

Apache Beam This tool is used to create a Hadoop based data pipeline model for streaming or batch processing that integrates with GCP. It is particularly useful for parallel processing and is suited to Extract, Transform, and Load (ETL) type tasks. So if we need to move data from one place to another, while performing transformations or calculations Beam is a good choice. And it will focus on the Python version to create a data pipeline [5]

The sample database stored in HDFS framework shown in Figure 4.

Company	3-Yr Stock Returns (%)	Net Sales (3-yr CAGR %)	Net Profit (3-yr CAGR %)
Dabur India	70.76	-0.15	4.91
Hindustan Unilever	93.86	2.20	3.68
Nestle India	99.74	6.67	23.33
Britannia Industries	125.34	4.80	7.83
Colgate-Palmolive (I)	51.82	1.01	3.79
Emami	-15.72	2.72	-8.75
Marico	38.15	2.02	7.27
Godrej Consumer Products	62.91	3.46	12.49
Jyothy Laboratories	29.94	3.01	8.97
Patanjali Ayurved	–	59.5	3.50

Compiled by: ETIG Database

Figure 4: FMCG stocks database

B. Data Injection: The Data was injected to HDFS based pipeline model. Pub/Sub is a messaging service that uses a Publisher-Subscriber model allowing us to ingest data in real-time.

Data Flow is a service that simplifies to create data pipelines and automatically handles the things like scaling up the infrastructure which can just concentrate on writing the code for our pipeline. The data injection will be needed to configure from local database to HDFS. This configuration has shown in Figure 5.

```
import csv
with open('FMCG.data', 'haircare') as csvfile:
        lines = csv.reader(csvfile)
        for row in lines:
                print ', '.join(row)
```

Figure 5: Configuring and splitting data

C. Hadoop framework is used by the following modules:

Hadoop Common: contains libraries and utilities needed by other Hadoop modules

Hadoop Distributed File System (HDFS): a distributed file-system that stores data on the commodity machines, providing very high aggregate bandwidth across the cluster

Hadoop YARN: a resource-management platform responsible for managing, compute resources in clusters and using them for scheduling of users' applications. [6]

Hadoop MapReduce: a programming model for large scale data processing

All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines, or racks of machines) are common and thus should be automatically handled in software by the framework. [6]

D. Pre-processing (Apache spark)

Apache Spark is a machine learning module. Apache Spark is a fast, in-memory data processing engine with elegant and expressive development APIs to allow data workers to efficiently execute streaming, machine learning or SQL workloads that require fast iterative access to datasets.[4] With Spark running on Apache Hadoop YARN, developers from everywhere can now create applications to exploit Spark's power, derive insights, and enrich their data science workloads within a single, shared dataset in Hadoop.

The Hadoop YARN-based architecture will be providing the foundation that enables Spark and other applications to share a common cluster and dataset while ensuring consistent levels of service and response. Spark is one of the many data access engines that works with YARN in HDP.

Loading dataset into spark engine using apache spark has been shown in Figure 6

```
import csv
import random
def loadDataset(filename, split, trainingSet=[] , testSet=[]):
        with open(filename, 'haircare') as csvfile:
            lines = csv.reader(csvfile)
            dataset = list(lines)
            for x in range(len(dataset)-1):
                for y in range(4):
                    dataset[x][y] = float(dataset[x][y])
                if random.random() < split:
                    trainingSet.append(dataset[x])
                else:
                    testSet.append(dataset[x])
```

Figure 6 : Apache spark program for inserting dataset in pyspark.

III. MACHINE LEARNING TECHNIQUES/ALGORITHMS

Classification is technique to categorize our data into a desired and distinct number of classes where we can assign label to each class. [7] [8]

Applications of Classification are: speech recognition, handwriting recognition, biometric identification, document classification etc.

Classifiers can be:

Binary classifiers: Classification with only 2 distinct classes or with 2 possible outcomes.

Multi-Class classifiers: Classification with more than two distinct classes.

K-NEAREST NEIGHBOUR (KNN):

KNN classified an object by a majority vote of the object's neighbours, in the space of input parameter. The object is assigned to the class which is most common among its k (an integer specified by human) nearest neighbour. It is a non-parametric, lazy algorithm. [8] Since, it's non-parametric, it does not make any assumption on data distribution (the data does not have to be normally distributed).

Because of the laziness, it does not really learn any model and make generalization of the data (It does not train some parameters of some function where input X gives output y). for these reasons, we can insist that this is not really a learning algorithm.

It simply classifies objects based on feature similarity (feature = input variables). [10] [11].

It has been shown in Figure 7 and 8

```
import math
def euclideanDistance(instance1, instance2, length):
            distance = 0
            for x in range(length):
            distance += pow((instance1[x] -
        instance2[x]), 2)
            return math.sqrt(distance)
```

Figure 7: classifying new data depending on Training instance distance

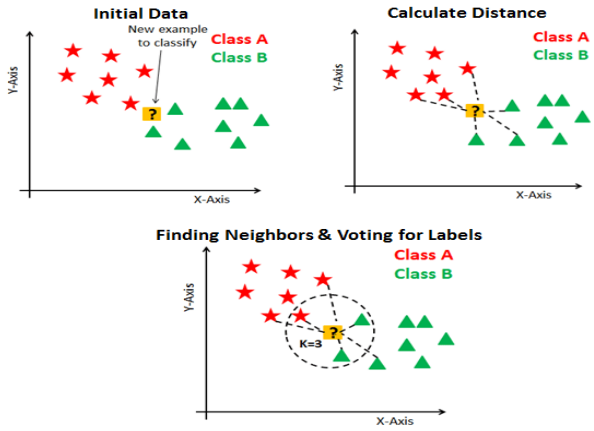


Figure 8: classifying and finding the featured data.

Algorithm: Implementing the KNN algorithm in Python

1. For each category of product Import the dataset from HDFS framework as CSV file.

2. Load the dataset and split our dataset into its

attributes and labels.

X = dataset.iloc[:, :-1].values

Y = dataset.iloc[:, 6].values

The X variable contains the first six columns of the dataset (i.e. attributes) while Y contains the labels.

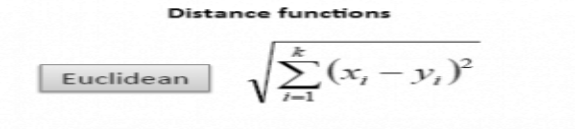
3. Train Test Split

To avoid over-fitting, we will divide our dataset into training and test splits, which presents us a better idea as to how our algorithm executed during the testing phase.

4. Discover the similar data from neighbour data.

Using feature extraction find similar data from neighbours.

To determine the similarity between two instances, we need a distance function. In our example, we are using the Euclidean distance formula.



Iterate the process till end of finding similar data sets from all the data

4. Perform Training data sets and Predictions.

Get response from neighbours by using

getResponse(neighbors):

classVotes = {}

Using get Accuracy function we will get the total correct predictions as a percentage of correct classifications. Classification is computed from a simple majority vote of the k nearest neighbours of each point.

Advantages: This algorithm is simple to implement, robust to noisy training data, and effective if training data is large.

Disadvantages: Need to determine the value of K and the computation cost is high as it needs to computer the distance of each instance to all the training samples.

IV.RESULTS AND DISCUSSION

The data is divided into training data and test data sets and make our linear regression-based learning model learn from the training data using KNN algorithm we can classify the trained data into training data sets between the different products attributes and values. Then the correlation between the sentiments and the FMCG market values is analysed based on the customer demand and feedback. Higher incomes aid growth in urban and rural markets.

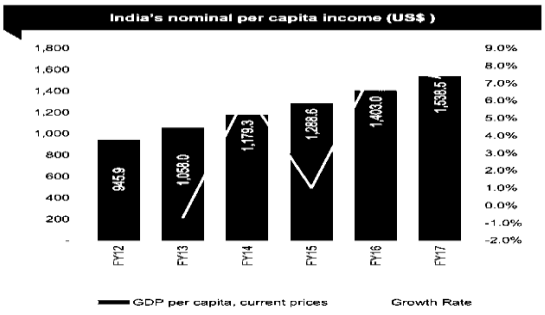


Figure 9: FMCG market income growth in India.

The learned model can be used to make future predictions about Market sales values.

FMCG is the 4th largest sector in the Indian economy.

Household and Personal Care is the leading segment, accounting for 50 per cent of the overall market. Hair care (23 per cent) and Food and Beverages (19 per cent) comes next in terms of market share, growing awareness, easier access and changing lifestyles have been the key growth drivers for the sector. The number of online users in India is likely to cross 850 million by 2025. Retail market in India is estimated to reach US$ 1.1 trillion by 2020 from US$ 672 billion in 2016, with modern trade expected to grow at 20 per cent - 25 per cent per annum, which is likely to boost revenues of FMCG companies People are gracefully embracing Ayurveda products, which has resulted in growth of FMCG major, Patanjali Ayurveda, with a m-cap of US$ 14.94 billion. The company s aims to expand globally in the next 5 to 10 years.

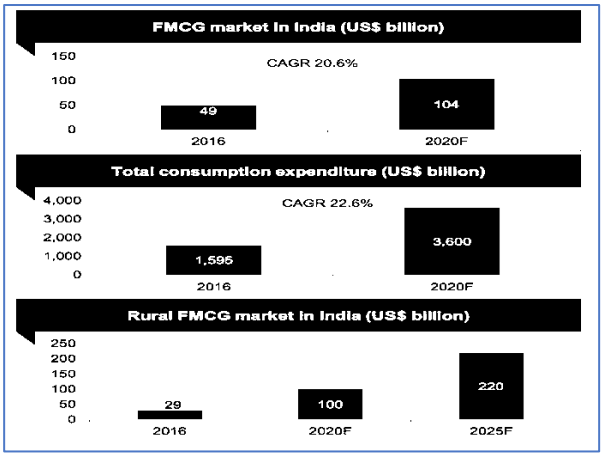


Figure 10: expecting sales in FMCG India market.

The growth in sales of major FMCG companies like Dabur, HUL, Marico, in the June-September 2017 quarter, is signalling the revival of consumer demand in India. Predictions from the test data shown in Figure 9 and 11.

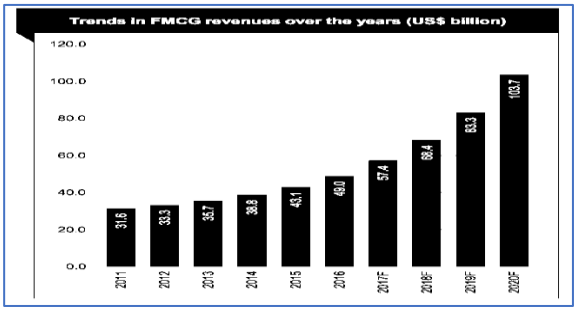


Figure: 11 FMCG income sales over the years.

Consumer products manufacturers ITC, Godrej Consumer Products Limited (GCPL) and HUL reported healthy net sales in FY17. [13]. Aggregate financial performance of the leading 10 FMCG companies over the past 8 quarters displays that the industry has grown at an average 16-21 per cent in the past 2 years.

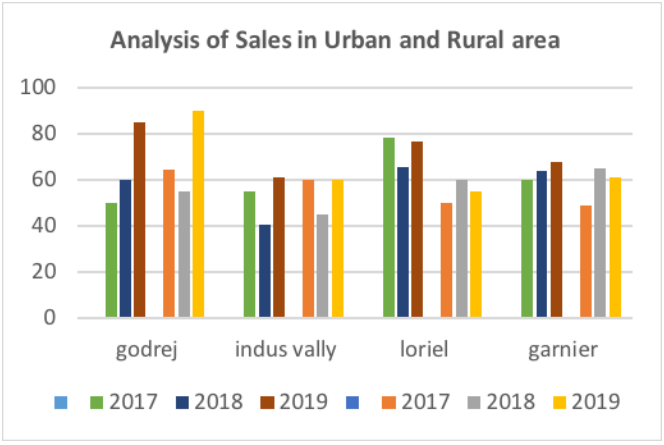


Figure 12: sale for selected companies in Rural and urban area.

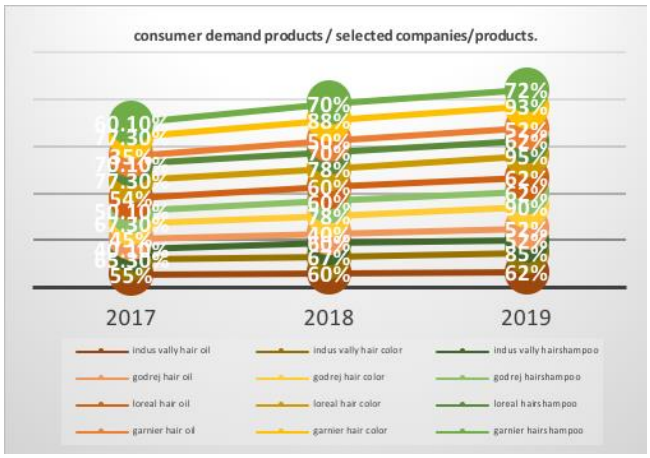


Figure 13: Analysis for demand fmcg goods for selected companies with selected products.

This analysis will help FMCG companies gain sales in Future. The big data analytics are true beneficial for finding the accurate predictions to increase the accuracy and correct analysis.

V. CONCLUSION and FUTURE WORKS

In this paper, the big data analytics are used for efficient FMCG market analysis and prediction. Through our work we were able to perform a predictive analysis on the selected category goods to help us identify the consumer demand FMCG products in urban and rural areas in everyday life. This will be helpful for FMCG market profit sales. With this Predictive analysis FMCG market will get the profitable trade in all aspects.

In the future enhancement, we can use this model by incorporating some additional features, such as greater focus on the healthier products, finding genuine quality and labelled products inspired by greater awareness. We also plan to test some Neural Network model-based learning rather than linear regression aims to accurately predict the FMCG market sales and marketing.

REFERENCES

1. http://www.ijaresm.com/uploaded_files/document_file/Jyotica_SinghkC6H.pdf

2. http://www.marketequations.com/services/retail-analytics-modeling-services.html

3. Stuart R. and Harald B.: Beginning Python for language Research, pp. 44 – 47, (2007).

4. Wiley:MachinelearninginPython,Predictivean. alysis, 2015.

5. FMCG (Fast Moving Consumer Goods) An Overview Jyotica Singh, (IJARESM) ISSN: 2455-6211, Volume 2, Issue 6, June-2014

6. Hadoop Map/Reduce tutorial. http://hadoop.apache.org/common/docs/r0.20.0/mapred tutorial.html.

7. Asha and Shravanthi, “Building Machine Learning Algorithms on Hadoop for Big Data”, in IJET UK Journal Vol 3, No 2 PP – 143 – 147, 2109.

8. https://www.dksh.com/hk-en/insights/seven-fmcg-trends-to-watch-out-for-2019

9. ApacheHive.http://hadoop.apache.org/hive.

10. Apche spark tutorial, https://www.toptal.com/spark/introduction-to-apache-spark

11. https://picnet.com.au/about-us/fmcg-machine-learning

12. https://simconblog.wordpress.com/2013/12/07/fmcg-sector-in-india/

13. https://www.ibef.org/states/andhra-pradesh.aspx#login-box

Mrs. Murari Thejovathi is a Research Scholar in the department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India. She is currently Assistant Professor in the department of Computer Science, King Khalid University, Abha, Kingdom of Saudi Arabia. Her research interests are Big Data Analytics and Security, Machine Learning.

Email: theju.scient@gmail.com

Dr. M.V.P. Chandra Sekhara Rao received his Ph.D. degree in Computer Science and Engineering from the Jawaharlal Nehru Technological University, Hyderabad. He is currently Professor in the Department of Computer Science and Engineering, RVR & JC College of Engineering, Guntur, Andhra Pradesh, India. His research interests are Data Mining, Big Data Analytics and Privacy Preserving in Data Mining.

Email: manukondach@gmail.com