A project report on

Machine Learning-driven Digital Marketing Insights

Submitted in partial fulfillment for the award of the degree of

Master of Science

In

Data Science

By

RETHANYA M

22MDT0122

Under the guidance of

Dr. KHADAR BABU SK

SCHOOL OF ADVANCED SCIENCES

VIT, Vellore

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

May 2024

CERTIFICATE

This is to certify that the thesis entitled **"MACHINE LEARNING–DRIVEN DIGITAL MARKETING INSIGHTS"** submitted by **RETHANYA M (Reg.No:22MDT0122) SCHOOL OF ADVANCED SCIENCES (SAS)**, VIT, for the award of the degree of MASTER OF SCIENCE in DATA SCIENCE, is a record of bonafide work carried out by him / her under my supervision during the period, **03.01.2024 to 07.05.2024**, as per the VIT code of academic and research ethics.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The thesis fulfills the requirements and regulations of the University and in my opinion meets the necessary standards for submission.

Place: Vellore
Date: 07.05.2024

07.05.24

Signature of the External Examiner

Signature of the Guide
Department of Mathematics
SAS, VIT-Vellore

9/5/2024

Head, Department of Mathematics
Dr. JAGADEESH KUMAR M.S
SAS, VIT- Vellore

DECLARATION

I hereby declare that the thesis entitled "**MACHINE LEARNING–DRIVEN DIGITAL MARKETING INSIGHTS**" submitted by me, for the award of the degree of Master of Science in Data Science to VIT is a record of Bonafide work carried out by me under the supervision of Dr. KHADAR BABU. SK

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute orany other institute or university.

Place: Vellore

Date: 07.05.2024

M. Rethy.

Signature of the Candidate

Rethanya M

ACKNOWLEDGEMENTS

With immense pleasure and deep sense of gratitude, I wish to express my sincere thanks to my guide Dr. KHADAR BABU SK, School of Advanced Sciences, VIT, Vellore without his motivation and continuous encouragement this research would not have been successfully completed.

I am grateful to the Chancellor of VIT, Vellore, Dr. G. Viswanathan, the Vice Presidents, and the Vice Chancellor for motivating me to carry out research in the Vellore Institute of Technology, Vellore and also for providing me with infrastructural facilities and many other resources needed for my research.

I express my sincere thanks to Dr. N Arunai Nambi Raj N, Dean, School of Advanced Sciences, VIT, Vellore, for his kind words of support and encouragement. I like to acknowledge the support rendered by my classmates in several ways throughout my research work.

I wish to thank Dr. JAGADEESH KUMAR M S, Head of Department, Mathematics, School of Advanced Sciences, VIT, Vellore for his encouragement and support.

I wish to extend my profound sense of gratitude to my parents and friends for all the support they made during my research and also providing me with encouragement whenever required.

M. Reth

Signature of the Student

Rethanya M

ABSTRACT

Now-a-days Online shopping involves direct customer transactions for goods and services. Real-time interaction with a seller over the internet, without the need for intermediaries. If a middleman is present, it is referred to as E-commerce, or electronic commerce. The objective of this initiative is to aims machine learning algorithms to anticipate sales of various products, including groceries, gadgets, and autos. Predicting sales is also known as sales forecasting. Estimating future sales involves anticipating how many products or services a salesperson, team, or firm will sell over a specific time period, such as a day, week, month, quarter, or year. Predicting sales in today’s market benefits both manufacturers and other companies who create parts. Use CNN to anticipate similar product recommendations based on customer evaluations and visually similar things found on websites.

Example: Product that are similar this.

The aim of this project is to prepare for online sales using machine learning. Machine learning algorithms are able to predict sales by identifying key patterns and variables inside data. Using an accurate forecasting model may improve food store profits and provide insights into better customer service. Also, forecast similar product recommendation systems.

.

TABLE OF CONTENTS

Page No:

ACKNOWLEDGEMENTS i

ABSTRACT 1

TABLE OF CONTENTS 2

LIST OF FIGURES 4

LIST OF TABLES 5

ABBREVIATIONS 5

1. INTRODUCTION 6

1.1 Objective 6

1.2 Motivation 7

1.3 Background 7

1.4 Literature Review 9

2. PROBLEM DESCRIPTION AND GOALS 13

2.1 Problem Statement 13

2.2 Project Goals 13

2.3 Methodology 14

3. TECHNICAL SPECIFICATION 18

3.1 Hardware Requirement 18

3.2 Software Requirement 18

4. DESIGN APPROACH AND DETAILS 21

4.1 Design Approach / Materials & Methods 21

4.2 Codes and Standards 23

5. SCHEDULE, TASKS AND MILESTONES 42

6. PROJECT OUTPUTS 43

7. ADVANTAGES & DISADVANTAGES 65

8. CONCLUSION 67

REFERENCES 68

LIST OF FIGURES

FIGURE NO	TITLE	PAGE NO
1.1	Methodology	14
1.2	Reading datasets from CSV files	16
1.3	Percentage of null values	44
1.4	Null data and drop the Null data	45
1.5	Summary of the product catalogue	45
1.6	Count product under each category	45
1.7	Top 10 selling products	46
1.8	Sales range of the product	47
1.9	Category wise average discount offered	49
1.10	Category wise BB and non-BB brands	50
1.11	Comparing old vs new recommenders	55
1.12	Correlation	56
1.13	Retrieve most similarities	62

LIST OF TABLES

TABLE NO	TITLE	PAGE NO
1.1	Schedule, Tasks and Milestone	42

ABBREVIATIONS

CNN- Convolutional neural network.

LSTM- Long Short term memory.

ML- Machine Learning.

DL- Deep Learning.

BB- Big Basket.

NLP- Natural language processing.

KPI- Key performance indicator.

KNN- K- nearest neighbor

1. INTRODUCTION

Today, Online Shopping is a process where customers connect with products, services and more. from seller to seller in real time. If there is an intermediary, it is called E-commerce or electronic commerce. The main aim of this project is to analyze food, furniture, vehicles, etc. using machine learning algorithms is to predict the sales of all kinds of product also called product forecasting. Sales forecasting predicts what products or services an individual salesperson, company will be able to sell over a specific time period, such as a day, week, month, quarter, or year. Estimating sales in the current market is not only for the industry but also for various other companies that produce parts. Also mention the similar products verification system using CNN based on customer reviews, which has similar products similar to those you see online.

1.1 OBJECTIVE

Online commerce has developed exponentially over the past decade and the industry is more dependent on online commerce than ever some time recently. It is fundamental to look at the diverse machines and the information mining handle to get it what the client likes and what they will be for the company. To assess the status of test in one of the online item categories. The investigate comes about with respect to the items to be sold so distant are wealthy and the inquire about strategies are assorted, the company has contracted numerous part-time and lasting representatives objective is for this extend to online deals utilizing machine learning. A great estimating framework can altogether increment a supermarket's benefit and is frequently advantageous to the organization as it increments productivity and gives understanding on how to way better serve clients. Suggestions for color decision-making exercises in common such as items to purchase, music to tune into online data to examined. The suggestion framework is exceptionally valuable when a contact needs to select from a number of potential benefit suppliers that can offer.

1.2 MOTIVATION

The sales forecasting to create way better techniques on educated forecasts past information is collected and analyzed through subjective models so that designs can be analyze, recognized and coordinate request arranging, future operations and showcasing operations. To estimate the deals and to anticipate whether the deals will increment by utilizing machine learning calculation., Proposal framework fundamental objective is to anticipate customer's interest and suggests the result things that are somewhat comparable to the costumer. Here we prescribe the comparable item by utilizing CNN, since CNNs do not require human supervision for the work of distinguishing vital highlights and fundamental preferences of CNNs is weight sharing. It moreover minimize the calculation in comparison with a normal neural organize. The primary preferences of this strategy is, it can rapidly construct the likeness between the modern data seen by comparable clients to each other .

1.3 BACKGROUND

` Predicting online sales involves using statistical and machine learning algorithms to forecast future sales of goods or services in an online retail environment. This prediction process relies on historical sales data, seasonal trends, marketing initiatives, customer behavior, and various other factors. Key variables impacting online sales predictions include time intervals (such as day, week, month, or year), marketing campaigns, customer behavior patterns, pricing strategies, competition dynamics, and machine learning techniques like regression, time-series forecasting, and neural networks.

Different time frames, including daily, weekly, monthly, or yearly intervals, exhibit varying sales patterns. For instance, sales might experience an uptick during peak seasons like holidays or significant events such as Black Friday.

- Marketing initiatives play a crucial role in boosting sales by enhancing brand visibility and promoting products through targeted ads or promotions.
- Understanding customer preferences and purchasing behavior, including demographics, browsing habits, and past purchases, enables businesses to forecast future sales accurately.

- Various pricing strategies, such as bundling, discounts, and dynamic pricing, can influence sales outcomes significantly.
- Competitor activities and market dynamics can also impact sales performance.
- Employing machine learning algorithms like regression, time-series forecasting, and neural networks can help analyze these factors and predict upcoming sales more effectively.
- Enhancing the model with additional data sources, optimizing parameters, and leveraging advanced techniques like deep learning and reinforcement learning can further enhance prediction accuracy. By leveraging these variables and advanced methods like deep learning and reinforcement learning, the accuracy of sales predictions can be enhanced.

In addition to sales prediction, similar product recommendation is a strategy used in e-commerce and online retail to suggest complementary products based on customer browsing and purchase history. This personalized recommendation approach aims to increase sales and customer loyalty. Recommendation systems utilize algorithms to analyze customer data, including demographics, purchase history, browsing behavior, and product reviews, to identify patterns and trends. These systems employ methods like collaborative filtering, content-based filtering, and hybrid systems to generate accurate and diverse product suggestions. By implementing these recommendation strategies, businesses can enhance the shopping experience, drive sales, and foster stronger customer relationships.

1.4 LITERATURE REVIEW

J Sekban, et al. (2019) proposed “Applying machine learning algorithms in sales prediction.” This thesis employs multiple distinct machine learning algorithms to achieve improved, optimal outcomes, which are further analyzed for forecasting purposes. Machine learning algorithms, like regression models and neural networks, are increasingly used for sales prediction. They analyze historical data, customer behavior, and market trends to forecast future sales accurately, aiding businesses in strategic planning and resource allocation. [1]

Behera, G., & Nain, N(2019) proposed a Grid Search Optimization (GSO) Based Future Sales Prediction For Big Mart, The authors of this study created a predictive model employing ensemble methodologies combined with the XGBoost Algorithm. Big marts are desiring most accurate forecasting techniques to ignore any losses on their investment. In forecasting future sales for Big Mart, statistical models like ARIMA and machine learning used algorithms are employed. By analyzing historical sales data, market trends, and seasonal variations, accurate predictions can be made, aiding in inventory management and strategic decision-making. [2]

Singh, B., Kumar, P., Sharma, N., & Sharma, K. P (2020) extensively explored and described a Sales Forecast for Amazon with Time series modeling, In B2C (business to consumer) e-commerce, precise deals forecasting plays a vital role in reducing costs and enhancing customer service experiences. This paper aims to analyze promotions related to unborn babies on Amazon.com without any copied content. It proposes three potential forecasting techniques—ARIMA, neural network auto-regression, and Holt Winters exponential smoothing—based on actual transaction data. Furthermore, it outlines accuracy metrics to evaluate the effectiveness of these forecasting methods using available sales data. The findings could assist Amazon in efficiently managing its forthcoming operations.[3]

S.K. Punjabi, S. Pranav. A. Yada, V. Shetty (2020) investigated the Sales prediction using Online Sentiment with Regression model. This essay is for to predict automobile sales by employing sentiment analysis sourced from various online platforms. The brand and online visibility of a vehicle are pivotal in its marketing strategy. Nevertheless, this article delves into several supplementary factors crucial for this analysis. Swift responses to market fluctuations benefit not only the manufacturer but also auxiliary industries involved in producing vehicle accessories or infrastructure. The essay employs polynomial regression models to project sales figures, specifically tailored to forecast sales in distinct regions.[4]

Singh, B., Kumar, P., Sharma, N., & Sharma, K. P (2020) Sales Forecast for Amazon Sales with Time Series Modeling, The main goal is to Forecasting accurately is crucial for minimizing expenses and enhancing customer service in B2C (Business to Consumer) e-commerce scenarios. The sales forecast for Amazon using time series modeling involves analyzing historical sales data to predict future sales accurately. This method considers various factors such as seasonality, trends, and external influences to generate reliable forecasts, essential for effective business planning and resource allocation. [5]

Purvika Bajaj, Renesa Ray, Shivani , ShravaniVidhate and et al,(2020) Sales Sales prediction using machine learning algorithms. The primary objective of this paper is to develop a forecasting dimension for future deals of major retail companies, with a focus on analyzing sales data from previous years. Sales prediction using machine learning algorithms involves analyzing historical sales data, market trends, and other relevant factors to forecast future sales accurately. By leveraging algorithms such as regression, decision trees, or neural networks, businesses can make data-driven decisions, optimize inventory, and enhance overall sales performance. [6]

Ranjitha P Spandana M (2021) introduced a Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms. The latest machine learning algorithm is highly sophisticated and offers techniques for predicting sales in various types of organizations, proving incredibly advantageous for mitigating the inexpensive methods used in forecasting. These are all for forecast sales trends, optimize inventory, and enhance decision-making. This data-driven approach leverages historical sales data, market variables, and customer behavior patterns to generate accurate predictions, empowering businesses to make informed strategies and boost profitability. [7]

T, G., Choudhary, R., & Prasad, et al,(2021) Prediction of Sales Value in online shopping using Linear Regression. The objective of this paper is to examine the transactions of a supermarket and improve the forecasting of their upcoming brands to align more effectively with market trends, thereby enhancing customer satisfaction.

Linear Regression involves analyzing historical data to identify trends and patterns that influence sales. This model uses statistical techniques to estimate future sales based on variables like customer demographics, purchase history, and marketing strategies, aiding businesses in making informed decisions. [8]

Sreemathy,J ; Prasath, et al,(2022) investigate a Machine Learning based Sales Prediction and Characterization using Consumer Perceptions. Machine Learning enables precise sales forecasts and consumer perception analysis. By leveraging ML algorithms on data from customer sentiments, buying patterns, and market trends, businesses can predict sales with accuracy and understand customer preferences for targeted marketing strategies, fostering growth and customer satisfaction. This sentence characterizes the model utilized for predicting demand based on historical data or trends.[9]

Nevil.K, Yashindev.H, et al, (2022) “Walmart sales forecasting using XGBoost algorithm and feature engineering” This paper introduces this sales forecasting model, which combines this algorithm with feature engineering processing to address Walmart's forecasting challenges in deals. This approach optimizes predictive accuracy by integrating data from multiple sources, including historical sales, market trends, and external factors. The result is a robust and efficient forecasting model tailored to dynamic retail environment. [10]

Yu Zheng, Quan Z. Sheng, and Li Sheng, (2023)"Machine Learning for Internet of Things Data Analysis. " published in IEEE Internet of Things Journal in 2018, provides an extensive overview of machine learning techniques applied to data analysis. The authors discuss various algorithms and frameworks used for extracting insights from the data, highlighting the importance of machine learning in deriving actionable intelligence from digital sources.[11]

Shahid N. Khan, Shahab Saquib Sohail, and Naeem Baig (2023) “Big Data Analytics in Healthcare: A Review on Machine Learning Techniques" published in Journal of Medical Systems in 2019, focuses on the application of machine learning in healthcare data analytics. The review discusses how machine learning algorithms such as deep learning and decision trees are utilized to extract valuable insights from large healthcare datasets, leading to improved diagnostics and patient care.[12]

Ashish K. Mishra, Suraj Sharma, and Deepak Garg, (2023) "Machine Learning in Financial Services: A Comprehensive Review" presents a comprehensive review of machine learning applications in the financial services sector. The authors explore how machine learning models are used for fraud detection, risk assessment, and investment prediction, emphasizing the role of data-driven insights in enhancing decision-making processes.[13]

Parul Gupta and Poonam Tanwar (2024) "Machine Learning for Social Media Analytics" published in Information Processing & Management in provides an in-depth survey of machine learning techniques for social media analytics. The review covers sentiment analysis, trend detection, and user behavior modeling using machine learning algorithms, showcasing how these methods contribute to generating actionable insights from digital social interactions.[14]

Xin Li, Jiaqing Tan, and et al, published the “Machine Learning-Driven Digital Marketing “, This paper focuses on the intersection of machine learning and digital marketing. The authors discuss how machine learning algorithms such as recommendation systems and customer segmentation models are revolutionizing digital marketing strategies by delivering personalized insights and enhancing customer engagement.[15]

2. PROBLEM DESCRIPTION AND GOALS

2.1 PROBLEM STATEMENT

Accurately predicting online sales proves challenging due to the intricate sales patterns, absence of reliable prediction models, and variable data quality. Factors such as unexpected events, shifts in consumer behavior, and competitive landscape changes significantly impact prediction model accuracy. Balancing precision with adaptability adds complexity, vital for businesses to excel in online sales, drive revenue, and satisfy customers. Efficient prediction must navigate variables like seasonality, market trends, and external influences, preventing inaccurate forecasts that lead to stock inefficiencies and increased costs. Accurate sales prediction empowers companies to optimize inventory levels, pricing strategies, and marketing campaigns, thereby improving overall operational efficiency. Similarly, personalized product recommendations aim to provide consumers with relevant suggestions based on their preferences and needs. Recommendation systems utilize customer information to anticipate their preferences and provide personalized product suggestions. Yet, establishing efficient recommendation strategies presents obstacles like algorithm intricacy, demands for high data quality, and the necessity to strike a balance between precision and variety to introduce customers to a diverse array of products.

2.2 PROJECT GOALS

1. From the datasets to predict the sales by using Machine Learning.
2. Identifying the top-selling prices
3. Finding sales price range of the product.
4. Finding products with high rating.
5. Count products under each category.
6. Increase supermarket future operations
7. Improve profits
8. Offering recommendations for similar products.
9. Finding accuracy by using models.

2.3 METHODOLOGY

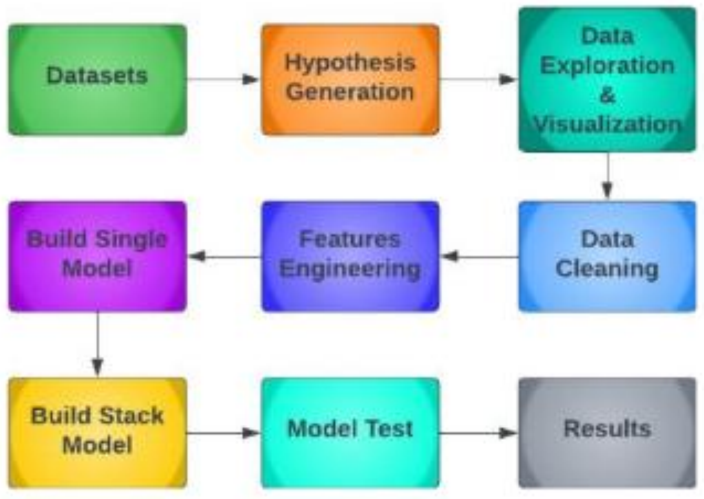


Figure 1.1

I. DATA COLLECTION

I gathered the datasets from the website kaggle.com. Specifically, I utilized test datasets comprising 5000 entries and train datasets with 8000 entries for this project.

II. DATA CLEANING

In typical data processing, we often encounter missing values and outliers. Our approach involves replacing missing values with the mode or mean of the corresponding trait, based on its nature, aiding in diminishing correlations among input features. Although managing outliers is vital in machine learning, advanced tree-based algorithms demonstrate robustness against outliers. Therefore, our focus lies on imputation techniques, a pivotal process for managing missing data and pinpointing outliers within the dataset.

III. HYPOTHESIS GENERATION

STORE LEVEL HYPOTHESIS:

- Stores with effective marketing strategies tend to experience sales growth due to enticing offers and appealing advertisements that attract customers.
- Organizing inventory based on local customer preferences by type and subcategory leads to increased sales.

PRODUCT LEVEL HYPOTHESIS:

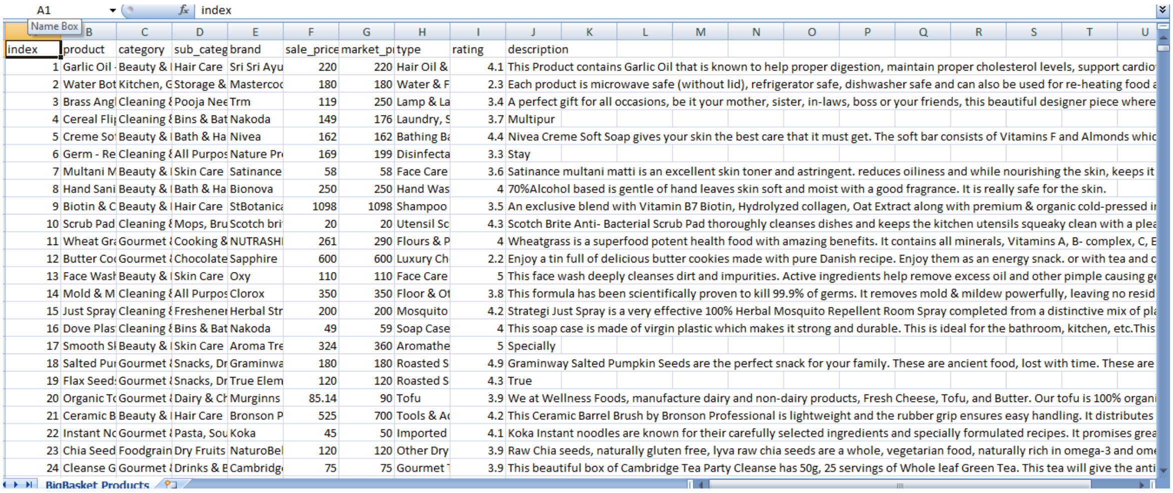
- **Brand**- Branded products generally achieve higher sales because of their perceived quality, quantity, and consumer trust.
- **Packaging**- Well-designed packaging without unnecessary additions can attract consumers and boost sales.
- **Utility**- Everyday use products are more likely to sell compared to specialized items. Products with high sales in stores tend to gain immediate attention and continue to sales.
- **Store Visibility-** Product placement within stores significantly influences sales; items near entrances are more noticed than those placed further back.
- **Advertising-** Increased in-store product advertising usually contributes to improved sales.

These are a few initial assertions I've outlined. Acknowledging that the existing data may not encompass every aspect, amalgamating these concepts facilitates a deeper comprehension of the matter and, when viable, promotes the exploration of further insights from publicly accessible outlets.

IV. DATA EXPLORATION

We will start by exploring the data and drawing conclusions based on our findings. Our goal is to identify any irregularities and discuss them in the next section.

The first step is to examine the data and compare it to our initial hypotheses. Below is a comparison between the data in the competition runner workbook and our assumptions.



A1 fx index

A	B	C	D	E	F	G	H	I	J
index	product	category	sub_categ	brand	sale_price	market_pi	type	rating	description
1	Garlic Oil -	Beauty & I	Hair Care	Sri Sri Ayu	220	220	Hair Oil &	4.1	This Product contains Garlic Oil that is known to help proper digestion, maintain proper cholesterol levels, support cardio
2	Water Bot	Kitchen, G	Storage &	Mastercoc	180	180	Water & F	2.3	Each product is microwave safe (without lid), refrigerator safe, dishwasher safe and can also be used for re-heating food a
3	Brass Angl	Cleaning &	Pooja Nee	Trm	119	250	Lamp & La	3.4	A perfect gift for all occasions, be it your mother, sister, in-laws, boss or your friends, this beautiful designer piece where
4	Cereal Flip	Cleaning &	Bins & Bat	Nakoda	149	176	Laundry, S	3.7	Multipur
5	Creme Sof	Beauty & I	Bath & Ha	Nivea	162	162	Bathing Ba	4.4	Nivea Creme Soft Soap gives your skin the best care that it must get. The soft bar consists of Vitamins F and Almonds whic
6	Germ - Re	Cleaning &	All Purpos	Nature Pr	169	199	Disinfecta	3.3	Stay
7	Multani M	Beauty & I	Skin Care	Satinance	58	58	Face Care	3.6	Satinance multani matti is an excellent skin toner and astringent. reduces oiliness and while nourishing the skin, keeps it
8	Hand Sani	Beauty & I	Bath & Ha	Bionova	250	250	Hand Was	4	70%Alcohol based is gentle of hand leaves skin soft and moist with a good fragrance. It is really safe for the skin.
9	Biotin & C	Beauty & I	Hair Care	StBotanica	1098	1098	Shampoo	3.5	An exclusive blend with Vitamin B7 Biotin, Hydrolyzed collagen, Oat Extract along with premium & organic cold-pressed ir
10	Scrub Pad	Cleaning &	Mops, Bru	Scotch bri	20	20	Utensil Sc	4.3	Scotch Brite Anti- Bacterial Scrub Pad thoroughly cleanses dishes and keeps the kitchen utensils squeaky clean with a plea
11	Wheat Gra	Gourmet &	Cooking &	NUTRASHI	261	290	Flours & P	4	Wheatgrass is a superfood potent health food with amazing benefits. It contains all minerals, Vitamins A, B- complex, C, E
12	Butter Coo	Gourmet &	Chocolate	Sapphire	600	600	Luxury Ch	2.2	Enjoy a tin full of delicious butter cookies made with pure Danish recipe. Enjoy them as an energy snack. or with tea and c
13	Face Wash	Beauty & I	Skin Care	Oxy	110	110	Face Care	5	This face wash deeply cleanses dirt and impurities. Active ingredients help remove excess oil and other pimple causing ge
14	Mold & M	Cleaning &	All Purpos	Clorox	350	350	Floor & Ot	3.8	This formula has been scientifically proven to kill 99.9% of germs. It removes mold & mildew powerfully, leaving no resid
15	Just Spray	Cleaning &	Freshener	Herbal Str	200	200	Mosquito	4.2	Strategi Just Spray is a very effective 100% Herbal Mosquito Repellent Room Spray completed from a distinctive mix of pla
16	Dove Plas	Cleaning &	Bins & Bat	Nakoda	49	59	Soap Case	4	This soap case is made of virgin plastic which makes it strong and durable. This is ideal for the bathroom, kitchen, etc.This
17	Smooth Sl	Beauty & I	Skin Care	Aroma Tre	324	360	Aromathe	5	Specially
18	Salted Pu	Gourmet &	Snacks, Dr	Graminwa	180	180	Roasted S	4.9	Graminway Salted Pumpkin Seeds are the perfect snack for your family. These are ancient food, lost with time. These are
19	Flax Seeds	Gourmet &	Snacks, Dr	True Elem	120	120	Roasted S	4.3	True
20	Organic To	Gourmet &	Dairy & Ch	Murginns	85.14	90	Tofu	3.9	We at Wellness Foods, manufacture dairy and non-dairy products, Fresh Cheese, Tofu, and Butter. Our tofu is 100% organi
21	Ceramic B	Beauty & I	Hair Care	Bronson P	525	700	Tools & Ac	4.2	This Ceramic Barrel Brush by Bronson Professional is lightweight and the rubber grip ensures easy handling. It distributes
22	Instant No	Gourmet &	Pasta, Sou	Koka	45	50	Imported	4.1	Koka Instant noodles are known for their carefully selected ingredients and specially formulated recipes. It promises grea
23	Chia Seed	Foodgrain	Dry Fruits	NaturoBel	120	120	Other Dry	3.9	Raw Chia seeds, naturally gluten free, lyva raw chia seeds are a whole, vegetarian food, naturally rich in omega-3 and ome
24	Cleanse G	Gourmet &	Drinks & B	Cambridge	75	75	Gourmet T	3.9	This beautiful box of Cambridge Tea Party Cleanse has 50g, 25 servings of Whole leaf Green Tea. This tea will give the anti

Figure 1.2

V. FEATURE ENGINEERING

Developing fresh input features for machine learning algorithms proves to be a potent strategy in crafting prediction models. Feature engineering encompasses the careful and alteration of variables to bolster prediction models, pinpointing vital data, identifying trends, and utilizing specialized knowledge. In the realm of forecast models, data usually comprises a target variable (the data slated for forecasting) alongside predictor variables. Once we delve into data exploration, we transition into tackling these variables and priming our data for analysis.

VI. MODELLING

Forecast modeling uses previously collected data to generate relevant results. Machine learning techniques are employed for this modeling. Select a model suitable for your data and the specific problem at hand; common models for prediction include linear regression, decision trees, and neural networks. Recommendation systems employ algorithms to assess customer information, such as demographic details, buying patterns, browsing habits, and product feedback, in order to detect recurring themes and emerging trends.

Evaluate the model's performance using metrics. Continuous monitoring and updating of the model are crucial to ensure accurate predictions. This might include gathering new data or adjusting model parameters as needed earlier insights.

VII. XGBOOST ALGORITHM

XG Boost stands out as a popular open-source rendition of the gradient boosted trees technique, a method in supervised learning that merges predictions from simpler models to forecast a target variable. Praised for its rapid processing, intuitive interface, and prowess with extensive datasets, XG Boost offers immediate usability post-installation, eliminating the need for parameter adjustments or fine-tuning.

VIII. CNN

CNNs are a distinct subset within deep learning methodologies, renowned for their prowess in tasks like image recognition and manipulation. They consist of multiple layers, such as convolutional, pooling, and fully connected layers. These layer filters identify features like edges, textures, and shapes in input images. These layers outputs pass through pooling layers, reducing spatial dimensions while preserving essential information. The final output from pooling layers is then fed into one or more fully connected layers, which predict and classify the image.

3. TECHNICAL SPECIFICATION

3.1 HARDWARE REQUIREMENT

The specifications for hardware outlined can form the basis of a contractual agreement for the system's implementation, presenting a thorough and uniform description of the complete system.

PROCESSOR:

Intel(R) Core(TM) i5

RAM : 4GB

HARD DISK: 443GB

These specifications delineate the essential hardware elements needed to operate the system.

3.2 SOFTWARE REQUIREMENT

The software requirements document functions as the architectural plan for the system, detailing the definitions and specifications of requirements. Its focus lies in delineating the system's intended achievements rather than the precise methodologies for reaching those objectives. Through a comprehensive breakdown of software requirements, it establishes the foundation for drafting the software requirements specification. This crucial document plays a pivotal role in cost estimation, team activity planning, task execution, and ongoing monitoring of the team's advancement during the development phase.

PYTHON IDE: Google – Colab Note Book

PROGRAMMING LANGUAGE: Python

SOFTWARE AND LANGUAGES USED

Tools and programming languages commonly employed in a standard Python machine learning workflow encompass a diverse selection that may include:

NumPy, utilized for manipulating matrices and vectors.

Pandas, specifically for managing time series and DataFrame structures to R.

Matplotlib, a library for creating 2D plots.

SciKit-Learn, serving as a repository for various machine learning algorithms and tools.

Tensor Flow, Powerful deep learning framework developed by Google, enabling efficient creation and deployment of complex neural networks.

Keras: High-level deep learning API simplifying neural network creation and training, compatible with back ends.

PYTHON

Python, an influential multi-functional programming language which boasts straightforward and user-friendly syntax rendering it ideal for beginners venturing into computer programming

Python's attributes include:

1. Coding

2. open-source nature

3. object-oriented design

4. high-level functionality

5. portability across platforms.

BUDGET

The project doesn't involve any physical components and is entirely executed using freely available software resources from the internet. Consequently, there's no designated budget assigned to this project.

4. DESIGN APPROACH AND DETAILS

4.1 Design Approach/ Materials & Methods

Here, we have used the following python libraries and methods for to get the results we want:

Numpy- This library is commonly utilized for executing numerical computations on arrays. In our project, we used this library to conduct numerical operations on the data frames extracted from the website. This enables us to compute moving averages for companies across multiple years**.**

Pandas- Pandas is commonly utilized for reading CSV files containing our products into a data frame, so, we can perform the execution of functions to achieve desired outcomes. Additionally, we leverage the Series module to extract specific columns from the data frame.

Math- We utilize the math module from Python to execute fundamental mathematical operations such as addition, subtraction, and multiplication on arrays since performing these operations directly on lists is not feasible.

Matplotlib - We utilized the widely used library to generate various graphs such as moving averages. Additionally, we imported the seaborn library from matplotlib to create graphs such as histograms, pair plots, and heat maps.

Cosine–Similarity - During sales prediction, we integrate cosine similarity to compare two items, this method to generate recommendations for similar products based on their ratings.

Arrays- We incorporate arrays in our processes to execute various digital marketing prediction calculations.

Vgg16- a convolutional neural network with multiple layers, specializes in object detection and classification, capable of identifying 1000 images across 1000 different categories.

ProgressBar- For our sales prediction model construction, we integrate a progress bar to visualize the batch-wise progress during model building.

Tensorflow - Here we used this library in Python to implement deep learning techniques for constructing our sales prediction model.

Keras- We imported this library for deep learning API, which operates on the TensorFlow machine learning library.

Sklearn - It stands out as a highly beneficial Python library for constructing machine learning models. Within our prediction model development project, we utilize functionalities like foundation of Machine Learning Evaluation and Data Preparation from this library.

4.2 CODES AND STANDARDS

CODING OF THE PROJECT

MOUNT THE GOOGLE DRIVE

```
from google.colab import drive

drive.mount('/content/drive')
```

IMPORT THE LIBRARIES

```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import numpy as np

import plotly.express as px

import plotly.graph_objects as go

from plotly.subplots import make_subplots

import colorama

from colorama import Fore, Back, Style

#ignoring warnings to keep the output clean

import warnings

warnings.filterwarnings('ignore')
```

READNG THE DATASETS

```
prod_data = pd.read_csv('/content/drive/MyDrive/kaggle/input/bigbasket-entire-product-list-28k-datapoints/BigBasket Products.csv (1)/BigBasket.csv')
```

CALCULATE THE DISCOUNT

```
prod_data["discount"] = (prod_data["market_price"] - prod_data["sale_price"]) /
prod_data["market_price"] * 100
prod_data["discount"]
```

NULL VALUES

```
#percentage of num values

print('Percentage Null Data In Each Column')
print('-'*30)
for col in prod_data.columns:
    null_count = prod_data[col].isnull().sum()
    total_count = prod_data.shape[0]
    print("{} : {:.2f}".format(col,null_count/total_count * 100))
#total NULL data
print('Total Null Data')
null_count = prod_data.isnull().sum().sum()
total_count = np.product(prod_data.shape)
print("{:.2f}".format(null_count/total_count * 100))
```

EXPLORATORY DATA ANALYSIS

SUMMARY OF THE DATASETS

```
print(Back.YELLOW+ Style.BRIGHT+ 'Summary of Product Catalogue:' +
Style.RESET_ALL)
print('Total number of unique ' + Fore.RED+ Style.BRIGHT+ 'Products' +
Style.RESET_ALL+'      :',\
      Fore.RED+ Style.BRIGHT+ str(len(prod_data['product'].unique())) +
Style.RESET_ALL)

print('Total number of ' + Fore.RED+ Style.BRIGHT+ 'Product Categories' +
Style.RESET_ALL +'   :',\
      Fore.RED+ Style.BRIGHT+ str(len(prod_data['category'].unique())) +
Style.RESET_ALL)

print('Total number of ' + Fore.RED+ Style.BRIGHT+ 'Sub-Categories' +
Style.RESET_ALL +'       :',\
    Fore.RED+ Style.BRIGHT+ str(len(prod_data['sub_category'].unique())) +
Style.RESET_ALL)

print('Total number of ' + Fore.RED+ Style.BRIGHT+ 'Product Types' +
Style.RESET_ALL +'       :',\
    Fore.RED+ Style.BRIGHT+ str(len(prod_data['type'].unique())) +
Style.RESET_ALL)

print('Total number of ' + Fore.RED+ Style.BRIGHT+ 'Brands' + Style.RESET_ALL
+'              :',\
    Fore.RED+ Style.BRIGHT+ str(len(prod_data['brand'].unique())) +
Style.RESET_ALL)
```

ANALYSIS OF CATEGORICAL FEATURES

```
#Count of Products under each Category
ctg_prod=prod_data[['category', 'product']]
ctg_prod=ctg_prod.drop_duplicates()
ctg_prod=ctg_prod.groupby('category').agg(prod_count=('product','count')).reset_inde
x().sort_values('prod_count', ascending=False)

# ctg_prod


fig = go.Figure(data=px.bar(x=ctg_prod.category,
                   y=ctg_prod.prod_count,
                   color = ctg_prod.category,
                   color_discrete_sequence=px.colors.sequential.Viridis  ,
                   title='<b>Count of Products under each Category</b>',
                   text = ctg_prod.prod_count,
                   height=500))
fig.update_layout(
   font_family="Times New Roman",
   title_font_family="Courier New",
   title_font_color="green",
   title_font_size=20,
   xaxis_title="<b>Category</b>",
   yaxis_title="<b>No. of Products</b>",
   legend_title_font_color="green"
)
```

TOP 10 SELLING PRODUCTS

```
#Top 10 selling Products
data = prod_data['product'].value_counts()[:10]
plt.figure(figsize=(10,8))
sns.barplot(x=data,y=data.index)
plt.xlabel('Count',fontdict={'fontsize': 25})
plt.ylabel('Product',fontdict={'fontsize': 25})
plt.title('Top 10 selling Products',fontweight="bold",fontdict={'fontsize': 30})
plt.rcParams['font.size'] = 10
```

ANALYSIS OF SALES PRICE

```
print(Back.YELLOW+ Style.BRIGHT+ 'Analysis of Sale Price:' +
Style.RESET_ALL)
print('Minimum Sale Price : '+ Fore.RED+ Style.BRIGHT+
str(prod_data['sale_price'].min()) + Style.RESET_ALL)
print('Maximum Sale Price : '+ Fore.RED+ Style.BRIGHT+
str(prod_data['sale_price'].max()) + Style.RESET_ALL)
print('Mean Sale Price    : '+ Fore.RED+ Style.BRIGHT+
str(round(prod_data['sale_price'].mean(),2)) + Style.RESET_ALL)
```

SALES PRICE RANGE OF THE PRODUCTS

```
#Sale Price Range of the Products
range_val = [['1-10',1, 10], ['11-25', 11, 25], ['26-50', 26, 50],
            ['51-100',51, 100], ['101-150', 101, 150], ['151-200', 151, 200],
            ['201-300',201, 300], ['301-400', 301, 400], ['401-500', 401, 500],
            ['501-1000',501, 1000], ['1001-1500', 1001, 1500], ['1501-2000', 1501, 2000],
            ['2001-3000',2001, 3000], ['3001-5000', 3001, 5000], ['5001-10000', 5001,
10000],
            ['10001-15000',10001, 15000]]
range_df =  pd.DataFrame(range_val, columns=['range_name', 'min_val', 'max_val'])
```

```
# range_df

range_df['prod_count']=''
for idx, row in range_df.iterrows():
    range_df.at[idx, 'prod_count'] = len(prod_data['product'][(prod_data['sale_price']>= row['min_val']) & (prod_data['sale_price']<= row['max_val'])])
# range_df

fig = go.Figure(data=px.bar(x=range_df.range_name,
                            y=range_df.prod_count,
                            color = range_df.range_name,
                            color_discrete_sequence=px.colors.sequential.Inferno,
                            title='<b>Sale Price Range of the Products</b>',
                            text = range_df.prod_count,
                            height=500))
fig.update_layout(
    font_family="Courier New",
    title_font_family="Courier New",
    title_font_color="red",
    title_font_size=20,
    xaxis_title="<b>Price Range</b>",
    yaxis_title="<b>No. of Products</b>",
    legend_title_font_color="green"
)
fig.show()
```

ANALYSIS OF DISCOUNTS

```
print(Back.YELLOW+ Style.BRIGHT+ 'Product with Highest Discount:' + Style.RESET_ALL)
prod_data[['product','category','discount']][prod_data['discount']==prod_data['discount'].max()]
```

TOP 20 DISCOUNTS

```
#Top 20 Disounts
print(Back.YELLOW+ Style.BRIGHT+ 'Top 20 Disounts:' + Style.RESET_ALL)
prod_data[['product','category','discount']].sort_values('discount',
ascending=False).head(20)
```

CATEGORY AND SUB-CATEGORY COUNT

```
#Category wise Average Discount Offered
catg_disc=prod_data[prod_data['discount']!=0].groupby('category').agg(avg_discount
=('discount','mean')).reset_index()
catg_disc=catg_disc.sort_values('avg_discount', ascending=False)
catg_disc= catg_disc.round({"avg_discount":2})

# catg_disc

fig = go.Figure(data=px.bar(x=catg_disc.category,
                    y=catg_disc.avg_discount,
                    color = catg_disc.category,
                    color_discrete_sequence=px.colors.sequential.Viridis,
                    title='<b>Category wise Average Discount Offered</b>',
                    text = catg_disc.avg_discount,
                    height=500))
fig.update_layout(
   font_family="Courier New",
   title_font_family="Courier New",
   title_font_color="Red",
   title_font_size=20,
   xaxis_title="<b>Category</b>",
   yaxis_title="<b>Average Discount</b>",
   legend_title_font_color="green"
)
fig.show()
```

BRAND AVERAGE RATING DETAILS

```
rating_df=prod_data[prod_data['rating'].notnull()].groupby('brand').agg(rating_count=
('rating', 'count'))\
     .reset_index().sort_values('rating_count', ascending=False)
rating_df= rating_df[rating_df['rating_count']>=50]

brand_avg_rating=prod_data[prod_data['brand'].isin(rating_df['brand'])].groupby(['br
and'])\
   .agg(avg_rating=('rating', 'mean')).reset_index().sort_values('avg_rating',
ascending=False)
brand_avg_rating= brand_avg_rating.round({"avg_rating":1})
brand_avg_rating=brand_avg_rating.sort_values('avg_rating', ascending=False)

# brand_avg_rating

print(Back.YELLOW+ Style.BRIGHT+ 'Brand Average Rating Details:' +
Style.RESET_ALL)
brand_avg_rating['avg_rating'].describe()
```

DEMOGRAPHIC FILTER RECOMMENDATION

```
#define
def sort_recommendor(col='rating',sort_type = False):
   """
   A recommendor based on sorting products on the column passed.
   Arguments to be passed:

   col: The Feature to be used for recommendation.
   sort_type: True for Ascending Order
   """
```

```
    rated_recommend = prod_data.copy()
    if rated_recommend[col].dtype == 'O':
        col='rating'
    rated_recommend = rated_recommend.sort_values(by=col,ascending = sort_type)
    return rated_recommend[['product','brand','sale_price','rating']].head(10)
```

SORTING

```
sort_recommendor(col='sale_price',sort_type=True)
C= prod_data['rating'].mean()
C
```

DEFINE

```
#define
def sort_recommendor(col='rating',sort_type = False):
    """
    A recommendor based on sorting products on the column passed.
    Arguments to be passed:

    col: The Feature to be used for recommendation.
    sort_type: True for Ascending Order
    """
    rated_recommend = prod_data.copy().loc[prod_data['rating'] >= 3.5]
    if rated_recommend[col].dtype == 'O':
        col='rating'
    rated_recommend = rated_recommend.sort_values(by=col,ascending = sort_type)
    return rated_recommend[['product','brand','sale_price','rating']].head(10)
```

CONTENT BASED RECOMMENDER

CONTENT-TEXT HANDLING LIBRARIES

```
#Text Handling Libraries
import re
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.metrics.pairwise import linear_kernel, cosine_similarity

tfidf = TfidfVectorizer(stop_words='english')
tfidf_matrix = tfidf.fit_transform(prod_data['description'])
tfidf_matrix.shape
```

COSINE SIMILARITY

```
cosine_sim = linear_kernel(tfidf_matrix, tfidf_matrix)
cosine_sim
```

GETTING RECOMMENDATION

```
indices = pd.Series(prod_data.index, index=prod_data['product']).drop_duplicates()

def get_recommendations_1(title, cosine_sim=cosine_sim):

    idx = indices[title]
    sim_scores = list(enumerate(cosine_sim[idx]))
    sim_scores = sorted(sim_scores, key=lambda x: x[1], reverse=True)
    sim_scores = sim_scores[1:11]
    movie_indices = [i[0] for i in sim_scores]
    return prod_data['product'].iloc[movie_indices]



get_recommendations_1('Water Bottle - Orange')
get_recommendations_1('Cadbury Perk - Chocolate Bar')
```

DATA COPY

```
prod_data1 = prod_data.copy()
prod_data1.head()
rmv_spc = lambda a:a.strip()
get_list = lambda a:list(map(rmv_spc,re.split('& |, |\*|\n', a)))

get_list('A & B, C')


for col in ['category', 'sub_category', 'type']:
    prod_data1[col] = prod_data1[col].apply(get_list)
```

TO AVOID DUPLICATE

```
def cleaner(x):
    if isinstance(x, list):
        return [str.lower(i.replace(" ", "")) for i in x]
    else:
        if isinstance(x, str):
            return str.lower(x.replace(" ", ""))
        else:
            return ''

for col in ['category', 'sub_category', 'type','brand']:
    prod_data1[col] = prod_data1[col].apply(cleaner)

def couple(x):
    return ' '.join(x['category']) + ' ' + ' '.join(x['sub_category']) + ' '+x['brand']+' ' +'
'.join( x['type'])
prod_data1['soup'] = prod_data1.apply(couple, axis=1)

prod_data1['soup'].head()
```

COUNT THE STRING VECTIORS

```
prod_data1.to_csv('data_cleaned_1.csv')

count = CountVectorizer(stop_words='english')
count_matrix = count.fit_transform(prod_data1['soup'])


cosine_sim2 = cosine_similarity(count_matrix, count_matrix)
cosine_sim2
```

COMPARING OLD AND NEW RECOMMENDATION

#CADBURY PERK

```
old_rec = get_recommendations_1('Cadbury Perk - Chocolate Bar').values
new_rec = get_recommendations_2('Cadbury Perk - Chocolate Bar',
cosine_sim2).values


pd.DataFrame({'Old Recommendor': old_rec,'New Recommendor':new_rec})
```

#WATER BOTTLE

```
old_rec = get_recommendations_1('Water Bottle - Orange').values
new_rec = get_recommendations_2('Water Bottle - Orange', cosine_sim2).values


pd.DataFrame({'Old Recommendor': old_rec,'New Recommendor':new_rec})
```

MODEL BUILDING

IMPORTING LIBRARIES

```
from sklearn.metrics import mean_squared_error, mean_absolute_error
from xgboost import XGBRegressor
from sklearn import ensemble
from lightgbm import LGBMRegressor
from sklearn.model_selection import cross_val_score
from catboost import CatBoostRegressor
```

ELIMINATING THE UNNECESSARY INDEXES

```
prod_data.drop("type",axis = 1,inplace=True)
prod_data.drop("is_bb_brand",axis = 1,inplace=True)
prod_data = prod_data.drop(prod_data[prod_data["market_price"]>1200].index)
prod_data.drop("brand",axis = 1,inplace = True)
prod_data.drop("product",axis = 1,inplace = True)
```

READING THE DATA

```
prod_data.head()
```

GETTING DUMMIES

```
from sklearn.preprocessing import MinMaxScaler

# Now you can use MinMaxScaler
scaler = MinMaxScaler()

X[column] = scaler.fit_transform(X[column])
print(X[column])
```

TRAINING THE DATASET

```
from sklearn.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=41)
```

XG BOOSTER ALGORITHM

```
xgb = XGBRegressor(booster='gbtree', objective='reg:squarederror')
from sklearn.model_selection import RandomizedSearchCV

param_lst = {
    "learning_rate" : [00.1,0.1,0.15,0.3,0.5],
    "n_estimators" : [100,500,1000,2000,3000],
    "max_depth" : [3,6,9],
    "min_child_weight" : [1,5,10,20],
    "reg_alpha" : [0.001,0.01,0.1],
    "reg_lambda" : [0.001,0.01,0.1]
}
xgb_reg = RandomizedSearchCV(estimator=xgb,param_distributions=param_lst,
                              n_iter = 5,scoring="neg_root_mean_squared_error",cv = 5)

xgb_reg = xgb_reg.fit(X_train,y_train)

best_param = xgb_reg.best_params_

xgb = XGBRegressor(**best_param)
xgb.fit(X_train,y_train)
preds = xgb.predict(X_test)
mae_xgb = mean_absolute_error(y_test,preds)
rmse_xgb = np.sqrt(mean_absolute_error(y_test,preds))
score_xgb = xgb.score(X_test,y_test)
cv_xgb = mean_cross_value(xgb,X,y)
model_performances = pd.DataFrame({
    "Model":["XGBoost"],
    "CV(5)" : [str(cv_xgb)],
    "MAE" : [str(mae_xgb)],
    "RMSE" : [str(rmse_xgb)],
    "Score" : [str(score_xgb)]
})
model_performances
```

PRODUCT RECOMMENDATION USING CNN

IMPORTING LIBRARIES

```


# Import libraries
import os
import cv2
import numpy as np
from PIL import Image
from sklearn.neighbors import NearestNeighbors

# Define parameters
data_dir = '/content/drive/My Drive/fashion-dataset'  # Path to the dataset
image_size = (128, 128)  # Size of the images

# imports
from keras.applications import vgg16
from tensorflow.keras.utils import load_img
from tensorflow.keras.utils import img_to_array
from keras.models import Model
from keras.applications.imagenet_utils import preprocess_input

from PIL import Image
import os
import matplotlib.pyplot as plt
import numpy as np
from sklearn.metrics.pairwise import cosine_similarity
```

```
import pandas as pd

import imageio
from keras.models import Model
from keras.applications import vgg16
import seaborn as sns
import matplotlib as mpl
import matplotlib.pyplot as plt
import matplotlib.image as mpimg
from PIL import Image, ImageOps, ImageFilter
import scipy.ndimage as ndi
from sklearn.metrics import classification_report, confusion_matrix

from keras.models import Sequential
from keras.layers import Dense, Flatten, Conv2D, MaxPool2D, Dropout, Activation, BatchNormalization
from keras.preprocessing.image import ImageDataGenerator
from keras.callbacks import ReduceLROnPlateau
from keras.preprocessing import image
from keras.utils import plot_model
```

PARAMETERS SETUP

```
imgs_path = "/content/drive/MyDrive/products/products/pro/style"
imgs_model_width, imgs_model_height = 224,224

nb_closest_images = 5 # number of most similar images to retrieve
```

LOAD THE VGG PRE-TRAINED MODEL FROM KERAS:

```
# load the model
vgg_model = vgg16.VGG16(weights='imagenet')


# remove the last layers in order to get features instead of predictions
```

```
feat_extractor = Model(inputs=vgg_model.input,
outputs=vgg_model.get_layer("fc2").output)

# print the layers of the CNN
feat_extractor.summary()
```

GETTING IMAGE PATH

```
import glob

# Define the directory containing the images
imgs_path = '/content/drive/MyDrive/products/products/pro/style'

# Find all the files with the extension `.png` in the directory
files = glob.glob(imgs_path + "/*.png")
```

NUMBER OF IMAGES:

```


# Print the number of images
print("number of images:", len(files))
```

FEED ONE IMAGE INTO CNN

```
!pip install pillow
from PIL import Image
# load an image in PIL format
original = load_img(files[0], target_size=(imgs_model_width, imgs_model_height))
plt.imshow(original)
plt.show()
print("image loaded successfully!")
```

GET THE EXTRACTED FEATURES

```
# get the extracted features
img_features = feat_extractor.predict(processed_image)

print("features successfully extracted!")
print("number of image features:",img_features.size)
img_features
```

FEEDING ALL IMAGES INTO CNN:

```
# load all the images and prepare them for feeding into the CNN

importedImages = []

for f in files:
    filename = f
    original = load_img(filename, target_size=(224, 224))
    numpy_image = img_to_array(original)
    image_batch = np.expand_dims(numpy_image, axis=0)

    importedImages.append(image_batch)
images = np.vstack(importedImages)
processed_imgs = preprocess_input(images.copy())
```

EXTRACTING ALL THE IMAGE FEATURES:

```
# extract the images features

imgs_features = feat_extractor.predict(processed_imgs)

print("features successfully extracted!")
imgs_features.shape
```

COMPUTE COSINE SIMILARITIES:

```
# compute cosine similarities between images

cosSimilarities = cosine_similarity(imgs_features)

# store the results into a pandas dataframe
cos_similarities_df = pd.DataFrame(cosSimilarities, columns=files, index=files)
cos_similarities_df.head()
```

RETRIEVING MOST SIMILARITIES:

```
def read_img(image_path):
    image = load_img(image_path,target_size=(224,224,3))
    image = img_to_array(image)
    image = image/222.
    return image
# function to retrieve the most similar products for a given one

def retrieve_most_similar_products(given_img):

    print("original product:")

    original = load_img(given_img, target_size=(imgs_model_width,
imgs_model_height))
    plt.imshow(original)
    plt.show()

    print("------------------------------------------------------------------------")
    print("most similar products:")

    closest_imgs =
cos_similarities_df[given_img].sort_values(ascending=False)[1:nb_closest_images+1
].index
```

```
    closest_imgs_scores =
cos_similarities_df[given_img].sort_values(ascending=False)[1:nb_closest_images+1
]

    for _ in range(1):
      i = random.randint(1,len(closest_imgs))
      plt.figure(figsize = (4 , 4))
      plt.figure(figsize = (20 , 20))

    for i in range(1,len(closest_imgs)):
        plt.subplot(1 , 5, i)
        plt.subplots_adjust(hspace = 0.5 , wspace = 0.3)
        original = load_img(closest_imgs[i], target_size=(imgs_model_width,
imgs_model_height))
        plt.imshow(original)
        plt.title(f'Similar Product #{i}')
        print("similarity score : ",closest_imgs_scores[i])
```

RETRIEVING SIMILAR PRODUCTS

```
retrieve_most_similar_products(files[2])
retrieve_most_similar_products(files[5])
retrieve_most_similar_products(files[3])
retrieve_most_similar_products(files[21])
retrieve_most_similar_products(files[39])
retrieve_most_similar_products(files[25])
```

5. SCHEDULE, TASKS AND MILESTONES

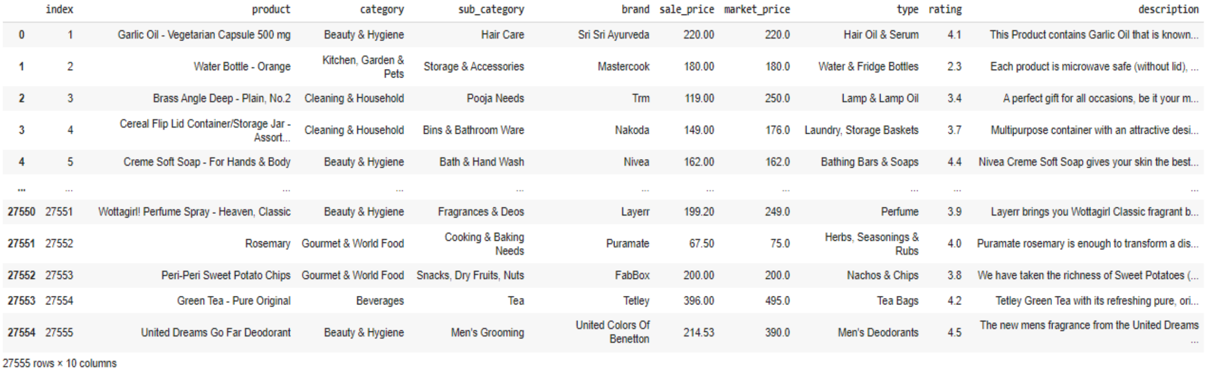
S.NO	MONTH-WEEK	PLAN
1.	JANUARY -WEEK 1	Identification of the problem.
2.	JANUARY – WEEK 2,3	Literature review on the decided problem
3.	JANUARY-WEEK 4	Discussion on the aims, objectives and outcomes of the problem.
4.	FEBRUARY-WEEK 1	Formation of abstract.
5.	FEBRUARY –WEEK 2	**C**ollection of data.
6.	FEBRUARY –WEEK 3	Methodology: Adaptation of the appropriate methods for the gathered data.
7.	FEBRUARY –WEEK 4	Appropriate analysis, relevant discussion and valid conclusions.
8.	MARCH –WEEK 1,2	Feedback from guide.
9.	MARCH –WEEK 3,4	Final documentation and Report writing.
10.	APRIL –WEEK 1-4	Report Review
11.	MAY-WEEK 1	Final Review

Table 1.1

6. PROJECT OUTPUTS

BIGBASKET:

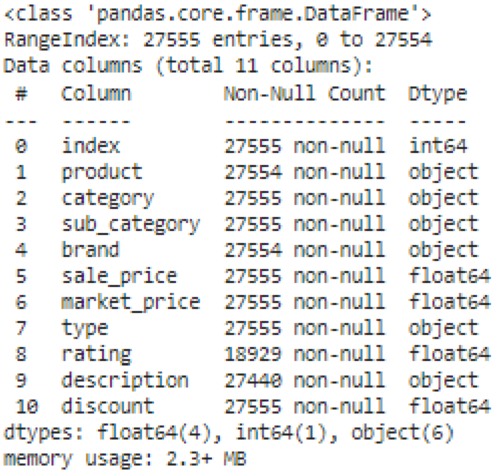
READING DATASETS FROM CSV FILES



	index	product	category	sub_category	brand	sale_price	market_price	type	rating	description
0	1	Garlic Oil - Vegetarian Capsule 500 mg	Beauty & Hygiene	Hair Care	Sri Sri Ayurveda	220.00	220.0	Hair Oil & Serum	4.1	This Product contains Garlic Oil that is known...
1	2	Water Bottle - Orange	Kitchen, Garden & Pets	Storage & Accessories	Mastercook	180.00	180.0	Water & Fridge Bottles	2.3	Each product is microwave safe (without lid), ...
2	3	Brass Angle Deep - Plain, No.2	Cleaning & Household	Pooja Needs	Trm	119.00	250.0	Lamp & Lamp Oil	3.4	A perfect gift for all occasions, be it your m...
3	4	Cereal Flip Lid Container/Storage Jar - Assort...	Cleaning & Household	Bins & Bathroom Ware	Nakoda	149.00	176.0	Laundry, Storage Baskets	3.7	Multipurpose container with an attractive desi...
4	5	Creme Soft Soap - For Hands & Body	Beauty & Hygiene	Bath & Hand Wash	Nivea	162.00	162.0	Bathing Bars & Soaps	4.4	Nivea Creme Soft Soap gives your skin the best...
...	...	...	...	...	...	...	...	...	...	...
27550	27551	Wottagirl! Perfume Spray - Heaven, Classic	Beauty & Hygiene	Fragrances & Deos	Layerr	199.20	249.0	Perfume	3.9	Layerr brings you Wottagirl Classic fragrant b...
27551	27552	Rosemary	Gourmet & World Food	Cooking & Baking Needs	Puramate	67.50	75.0	Herbs, Seasonings & Rubs	4.0	Puramate rosemary is enough to transform a dis...
27552	27553	Peri-Peri Sweet Potato Chips	Gourmet & World Food	Snacks, Dry Fruits, Nuts	FabBox	200.00	200.0	Nachos & Chips	3.8	We have taken the richness of Sweet Potatoes (...
27553	27554	Green Tea - Pure Original	Beverages	Tea	Tetley	396.00	495.0	Tea Bags	4.2	Tetley Green Tea with its refreshing pure, ori...
27554	27555	United Dreams Go Far Deodorant	Beauty & Hygiene	Men's Grooming	United Colors Of Benetton	214.53	390.0	Men's Deodorants	4.5	The new mens fragrance from the United Dreams ...

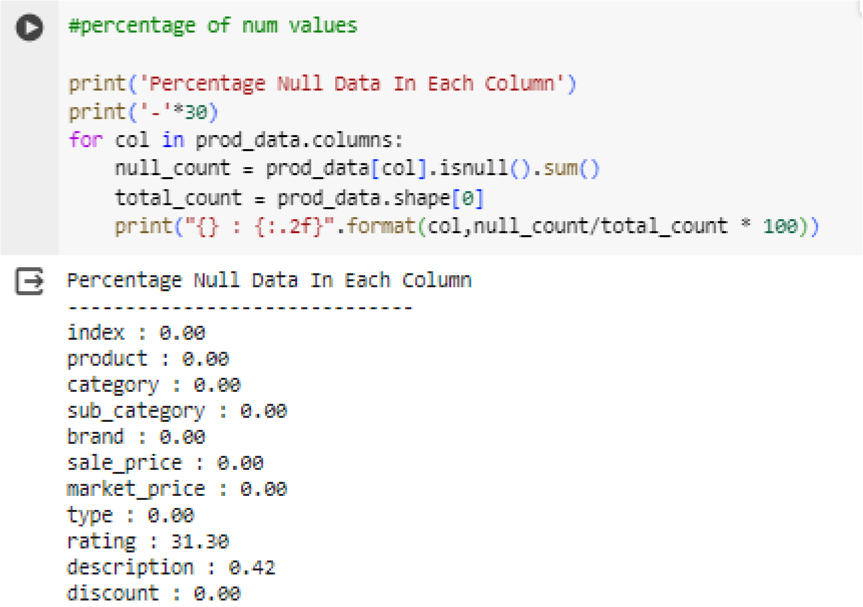
27555 rows × 10 columns

DATA INFORMATION



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27555 entries, 0 to 27554
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   index         27555 non-null  int64  
 1   product       27554 non-null  object 
 2   category      27555 non-null  object 
 3   sub_category  27555 non-null  object 
 4   brand         27554 non-null  object 
 5   sale_price    27555 non-null  float64
 6   market_price  27555 non-null  float64
 7   type          27555 non-null  object 
 8   rating        18929 non-null  float64
 9   description   27440 non-null  object 
 10  discount      27555 non-null  float64
dtypes: float64(4), int64(1), object(6)
memory usage: 2.3+ MB
```

PERCENTAGE OF NULL VALUES IN EACH COLUMNS

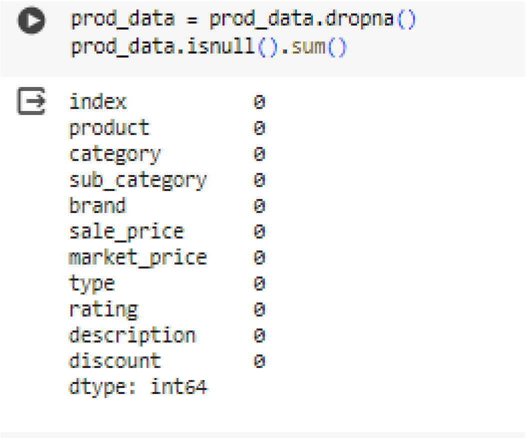


```
#percentage of num values

print('Percentage Null Data In Each Column')
print('-'*30)
for col in prod_data.columns:
    null_count = prod_data[col].isnull().sum()
    total_count = prod_data.shape[0]
    print("{} : {:.2f}".format(col,null_count/total_count * 100))
```

```
Percentage Null Data In Each Column
------------------------------
index : 0.00
product : 0.00
category : 0.00
sub_category : 0.00
brand : 0.00
sale_price : 0.00
market_price : 0.00
type : 0.00
rating : 31.30
description : 0.42
discount : 0.00
```

Figure 1.3

DROP THE NULL DATA

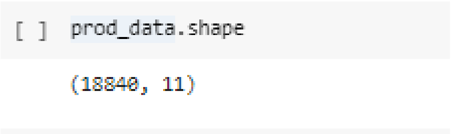


```
prod_data = prod_data.dropna()
prod_data.isnull().sum()
```

```
index           0
product         0
category        0
sub_category    0
brand           0
sale_price      0
market_price    0
type            0
rating          0
description     0
discount        0
dtype: int64
```

Figure 1.4

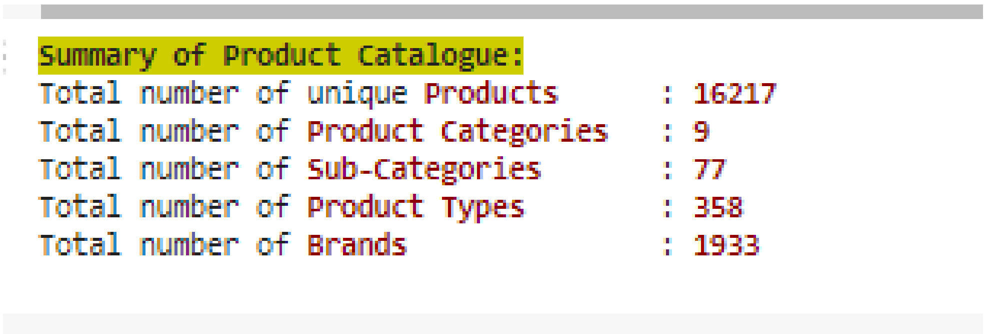
PRODUCT SHAPE AFTER DROPPED NULL VALUES:



```
[ ]  prod_data.shape

     (18840, 11)
```

EXPLORATORY DATA ANALYSIS:

SUMMARY OF PRODUCT CATALOGUE



```
Summary of Product Catalogue:
Total number of unique Products      : 16217
Total number of Product Categories   : 9
Total number of Sub-Categories       : 77
Total number of Product Types        : 358
Total number of Brands               : 1933
```

Figure 1.5

ANALYSIS OF CATEGORICAL FEATURES:

PRODUCT, PRODUCTTYPE,CATEGORYANDSUB-CATEGORY:

COUNT OF PRODUCTS UNDER EACH CATEGORY

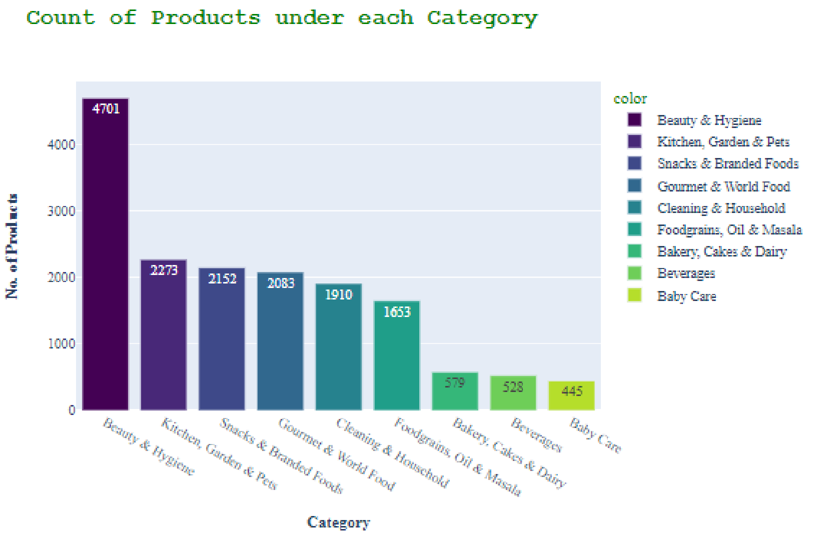


Figure 1.6

OBSERVATION: In category Beauty &hygiene has the highest number of products which is followed by Kitchen , Garden & Pets and Gournet & World Food.

TOP 10 SELLING PRODUCTS

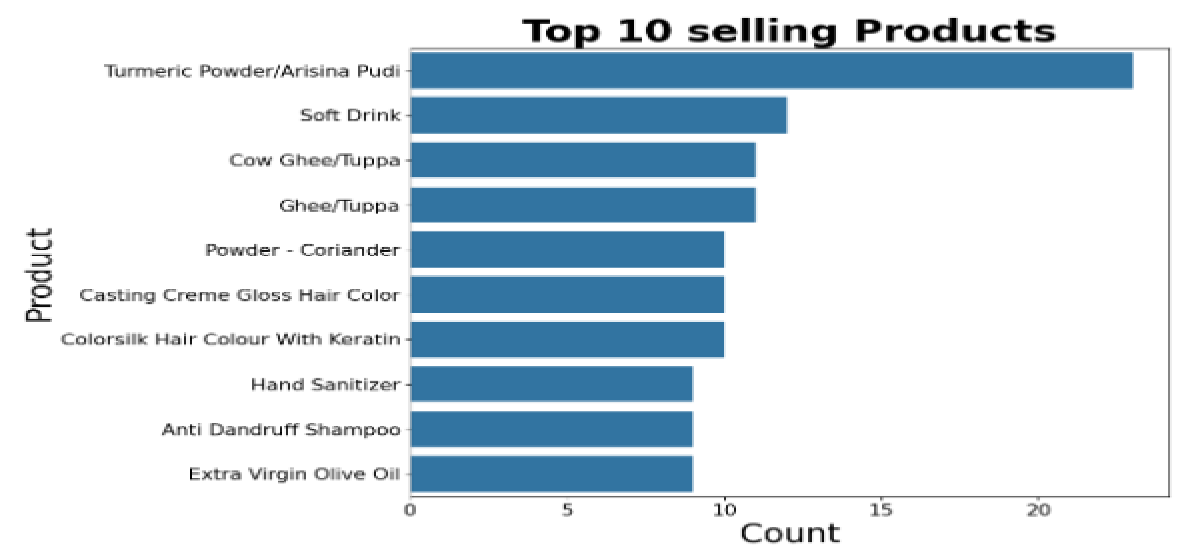
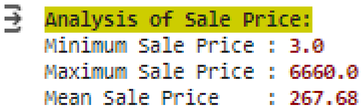


Figure 1.7

OBSERVATION: In Brand, Fresho has the highest number of Product Types, followed by bb Royal and BB Home

ANALYSIS OF SALES PRICE



```
Analysis of Sale Price:
Minimum Sale Price : 3.0
Maximum Sale Price : 6660.0
Mean Sale Price    : 267.68
```

OBSERVATION: With minimum price at 3 and maximum price is at 6666, range appears to be huge but mean is only 267.

SALE PRICE RANGE OF THE PRODUCTS

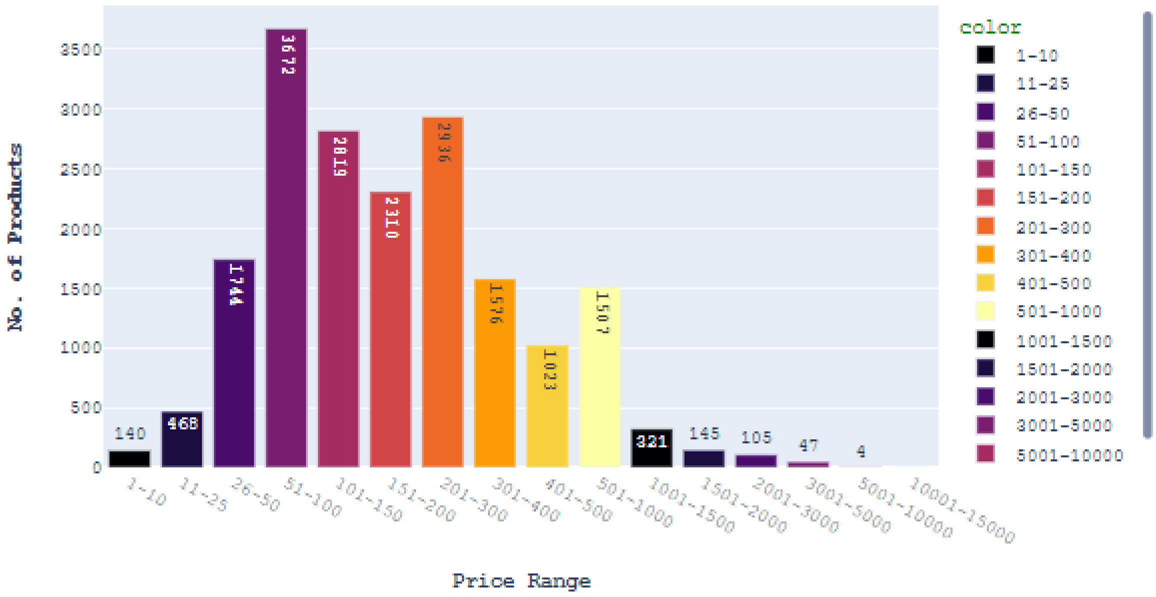
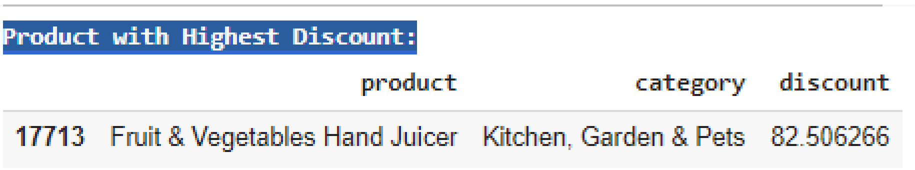


Figure 1.8

OBSERVATION: The data from the plot indicates that 28% of the products are priced below Rs100, 53% fall within the range of 1-200, 78% are within 1-400, and 95% are within 1-1000.

ANALYSIS OF DISCOUNTS

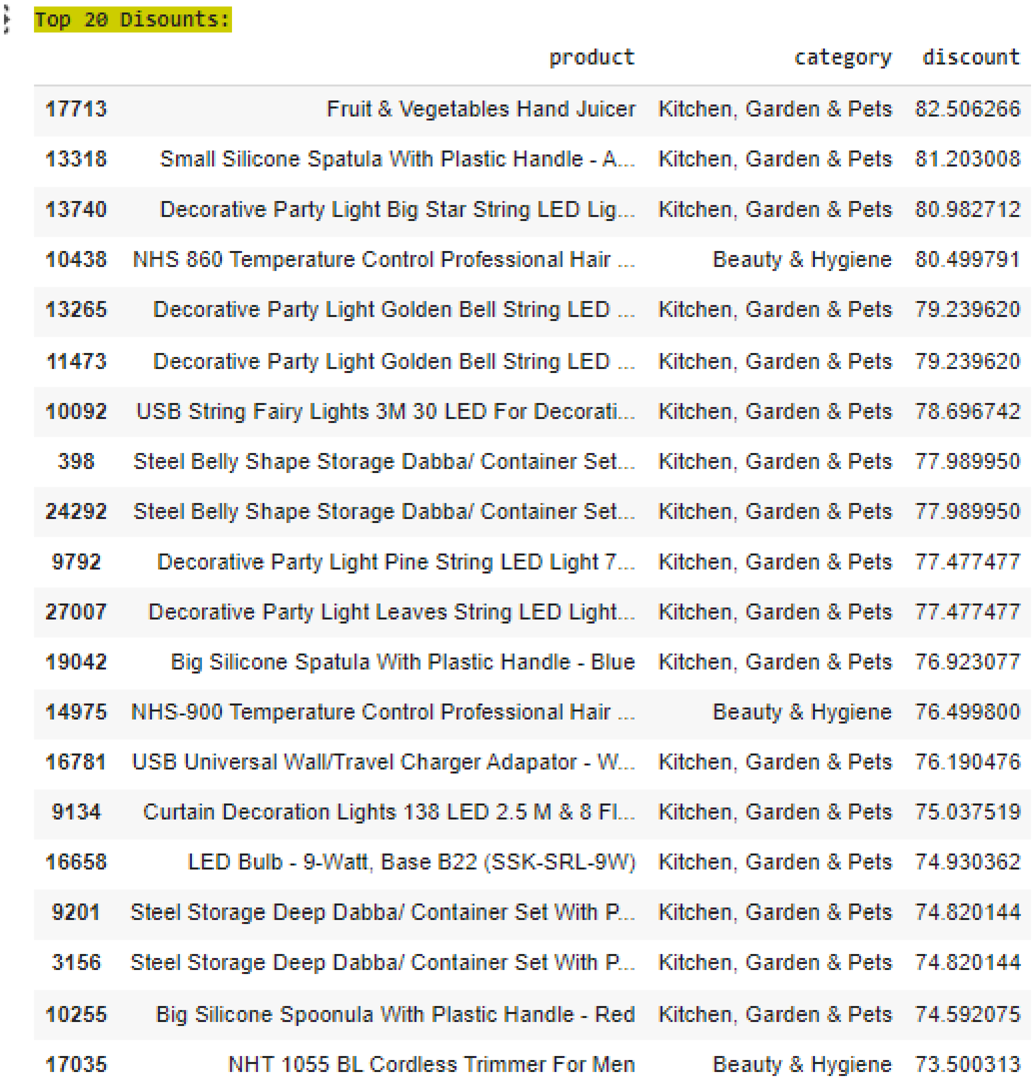
PRODUCT WITH HIGHEST DISCOUNT



Product with Highest Discount:

	product	category	discount
17713	Fruit & Vegetables Hand Juicer	Kitchen, Garden & Pets	82.506266

TOP 20 DISCOUNTS



```
Top 20 Disounts:
```

	product	category	discount
17713	Fruit & Vegetables Hand Juicer	Kitchen, Garden & Pets	82.506266
13318	Small Silicone Spatula With Plastic Handle - A...	Kitchen, Garden & Pets	81.203008
13740	Decorative Party Light Big Star String LED Lig...	Kitchen, Garden & Pets	80.982712
10438	NHS 860 Temperature Control Professional Hair ...	Beauty & Hygiene	80.499791
13265	Decorative Party Light Golden Bell String LED ...	Kitchen, Garden & Pets	79.239620
11473	Decorative Party Light Golden Bell String LED ...	Kitchen, Garden & Pets	79.239620
10092	USB String Fairy Lights 3M 30 LED For Decorati...	Kitchen, Garden & Pets	78.696742
398	Steel Belly Shape Storage Dabba/ Container Set...	Kitchen, Garden & Pets	77.989950
24292	Steel Belly Shape Storage Dabba/ Container Set...	Kitchen, Garden & Pets	77.989950
9792	Decorative Party Light Pine String LED Light 7...	Kitchen, Garden & Pets	77.477477
27007	Decorative Party Light Leaves String LED Light...	Kitchen, Garden & Pets	77.477477
19042	Big Silicone Spatula With Plastic Handle - Blue	Kitchen, Garden & Pets	76.923077
14975	NHS-900 Temperature Control Professional Hair ...	Beauty & Hygiene	76.499800
16781	USB Universal Wall/Travel Charger Adaptor - W...	Kitchen, Garden & Pets	76.190476
9134	Curtain Decoration Lights 138 LED 2.5 M & 8 Fl...	Kitchen, Garden & Pets	75.037519
16658	LED Bulb - 9-Watt, Base B22 (SSK-SRL-9W)	Kitchen, Garden & Pets	74.930362
9201	Steel Storage Deep Dabba/ Container Set With P...	Kitchen, Garden & Pets	74.820144
3156	Steel Storage Deep Dabba/ Container Set With P...	Kitchen, Garden & Pets	74.820144
10255	Big Silicone Spoonula With Plastic Handle - Red	Kitchen, Garden & Pets	74.592075
17035	NHT 1055 BL Cordless Trimmer For Men	Beauty & Hygiene	73.500313

CATEGORY WISE AVERAGE DISCOUNT OFFERED

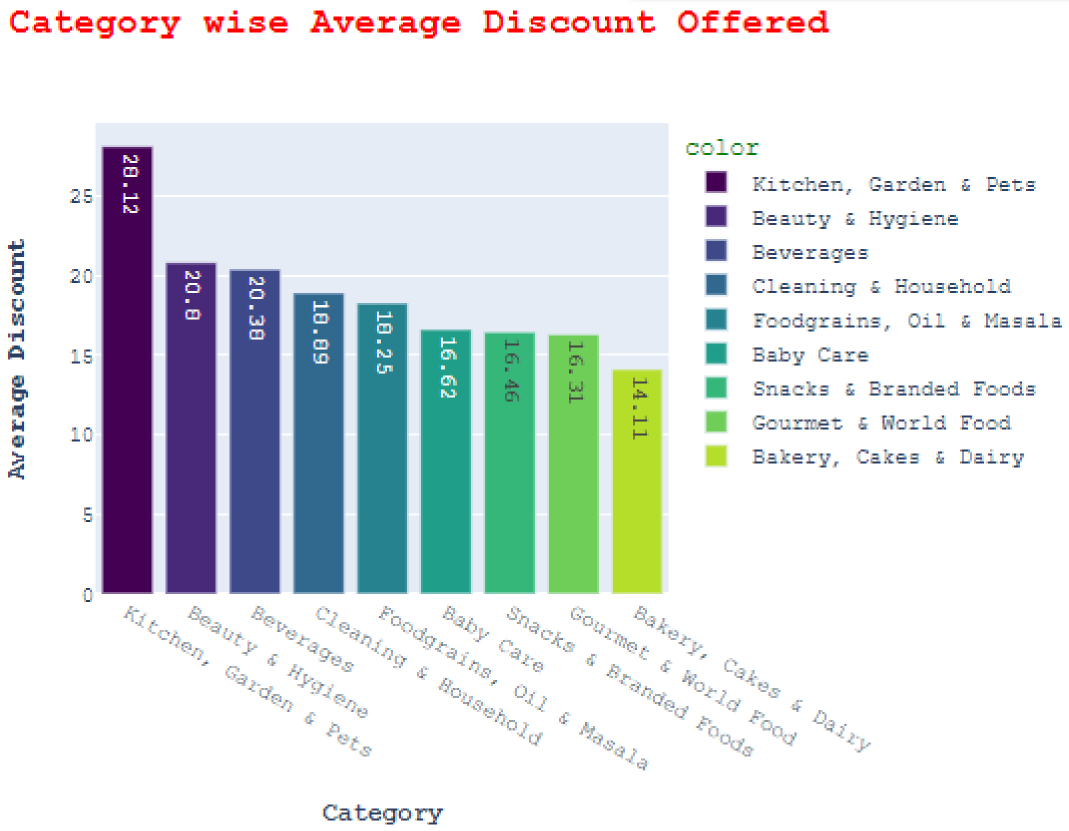


Figure 1.9

OBSERVATION: The Highest average discount is observed in kitchen, Garden & pets category, followed by Fruits, Beverages, Beauty & hygiene, Vegetables & baby care products.

CATEGORY WISE AND NON-BB BRANDS

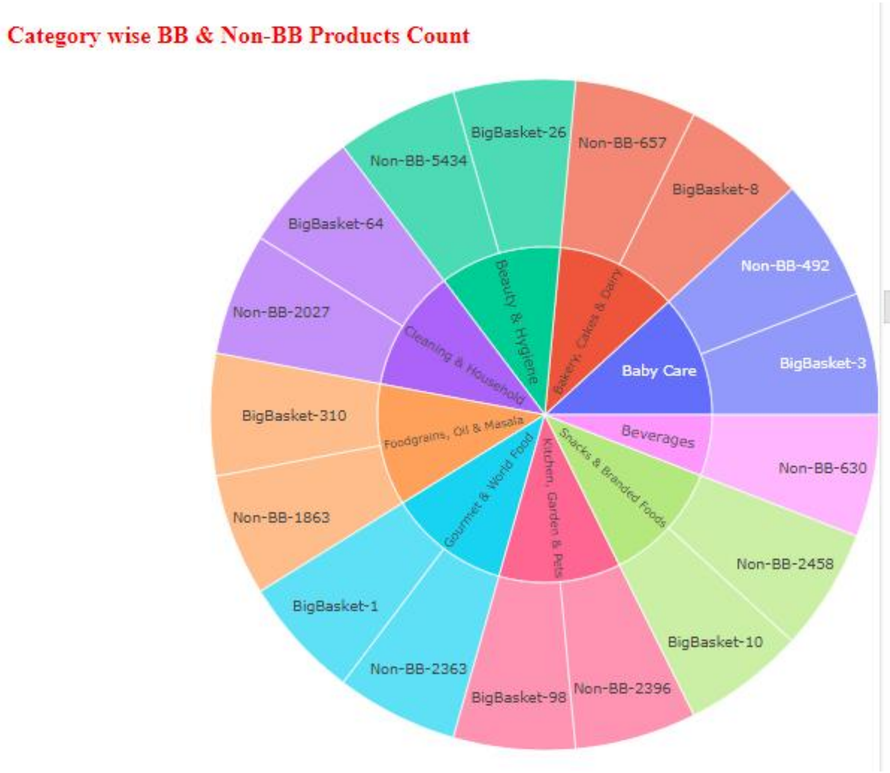


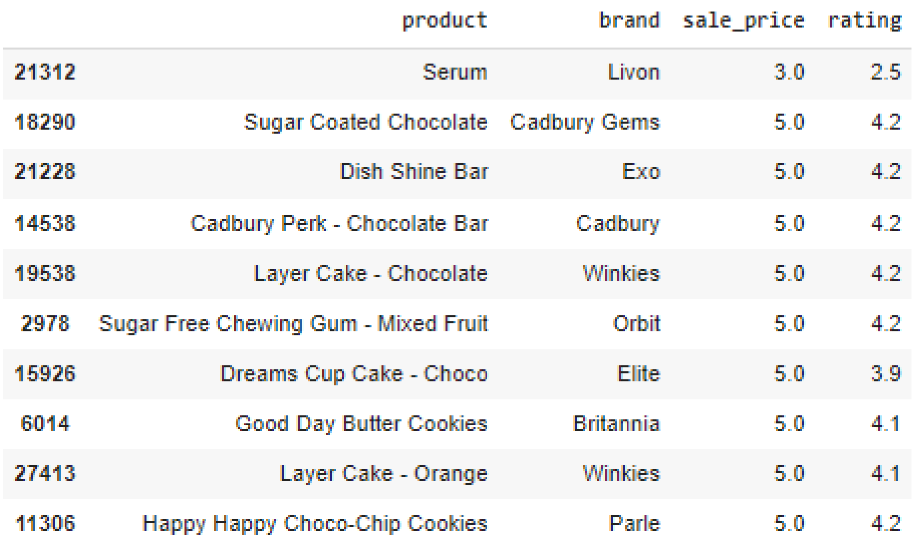
Figure 1.10

Big Basket's own brands dominate the Fruits & Vegetables category, accounting for 97% of the products sold. In the Eggs, Meat & Fish category, as well as the Food Grains, Oil & Masala Category, Big Basket brands hold market shares of 31% and 23%, respectively. This data underscores the substantial presence of Big Basket brands in the household's everyday cooking essentials. Among Big Basket's brands, Fresho, BB Royal, and BB Home offer the most diverse range of product types. However, Big Basket's brands have minimal representation, less than 1%, in categories such as Baby Care, Beauty & Hygiene, and Snacks & Branded Foods.

DEMOGRAPHIC FILTER RECOMMENDATION

Demographic Filtering is like Suggesting items based on a characteristic, such as the top 10 rated items or the top 10 items within a specific category**.**

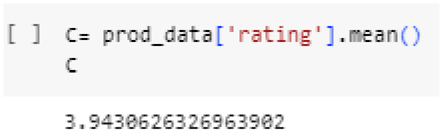
SORTING RECOMMEND



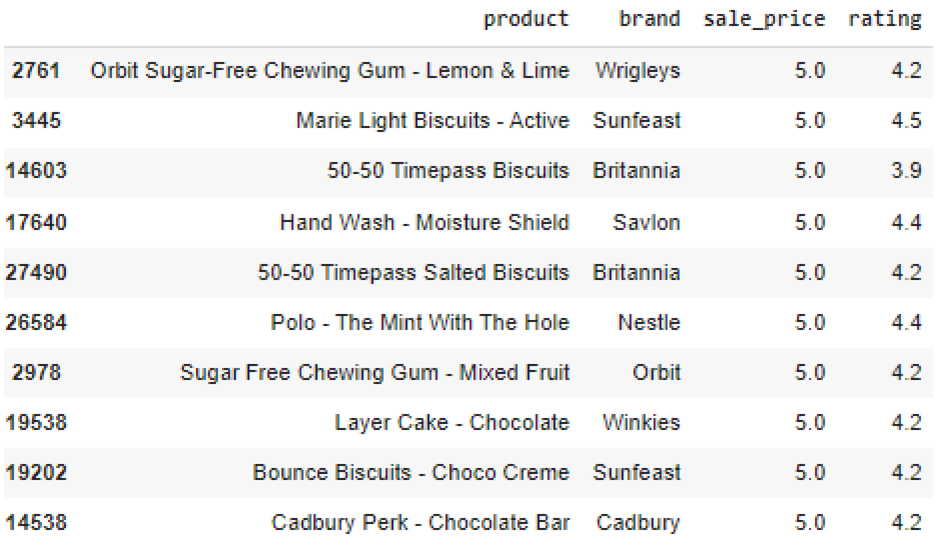
	product	brand	sale_price	rating
21312	Serum	Livon	3.0	2.5
18290	Sugar Coated Chocolate	Cadbury Gems	5.0	4.2
21228	Dish Shine Bar	Exo	5.0	4.2
14538	Cadbury Perk - Chocolate Bar	Cadbury	5.0	4.2
19538	Layer Cake - Chocolate	Winkies	5.0	4.2
2978	Sugar Free Chewing Gum - Mixed Fruit	Orbit	5.0	4.2
15926	Dreams Cup Cake - Choco	Elite	5.0	3.9
6014	Good Day Butter Cookies	Britannia	5.0	4.1
27413	Layer Cake - Orange	Winkies	5.0	4.1
11306	Happy Happy Choco-Chip Cookies	Parle	5.0	4.2

OBSERVATION: Here we can see that our top product's rating has 2.5 which is quite bad, So let us filter it down by setting a threshold rating

RATING



```
[ ]  C= prod_data['rating'].mean()
     C

     3.9430626326963902
```



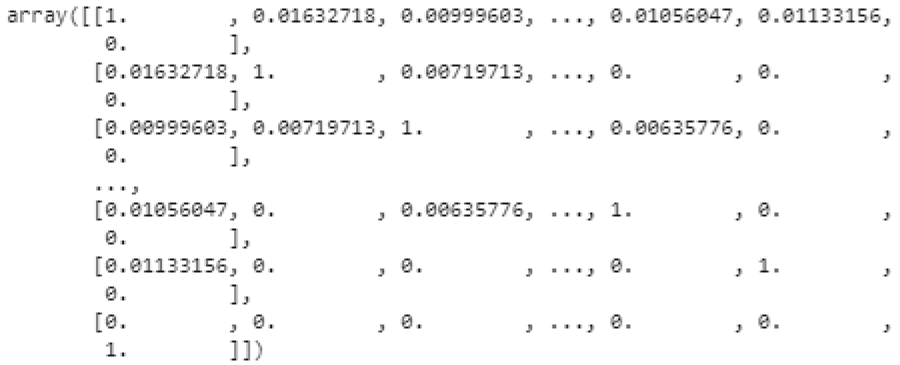
	product	brand	sale_price	rating
2761	Orbit Sugar-Free Chewing Gum - Lemon & Lime	Wrigleys	5.0	4.2
3445	Marie Light Biscuits - Active	Sunfeast	5.0	4.5
14603	50-50 Timepass Biscuits	Britannia	5.0	3.9
17640	Hand Wash - Moisture Shield	Savlon	5.0	4.4
27490	50-50 Timepass Salted Biscuits	Britannia	5.0	4.2
26584	Polo - The Mint With The Hole	Nestle	5.0	4.4
2978	Sugar Free Chewing Gum - Mixed Fruit	Orbit	5.0	4.2
19538	Layer Cake - Chocolate	Winkies	5.0	4.2
19202	Bounce Biscuits - Choco Creme	Sunfeast	5.0	4.2
14538	Cadbury Perk - Chocolate Bar	Cadbury	5.0	4.2

Take note that the product rate is at 2.5 is currently not recommended. This was our initial suggestion, known for its simplicity, effectiveness.

CONTENT BASED RECOMMENDER

To calculate the similarity score, we will import the Linear_Kernel, which computes the dot product of the tfidf_matrix and provides an aggregate value representing the similarity score.

COSINE SIMILARITY

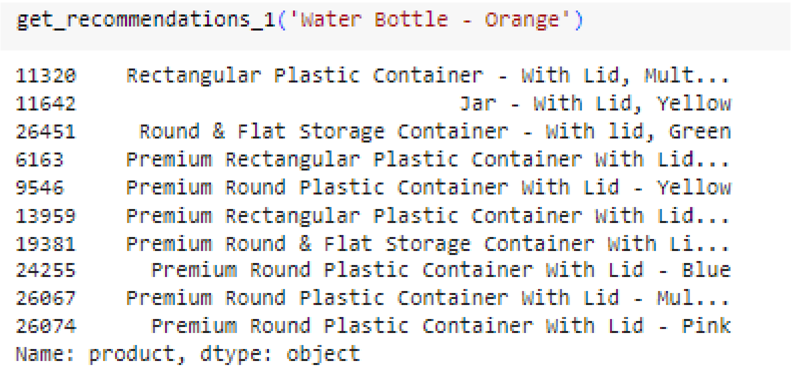


```
array([[1.        , 0.01632718, 0.00999603, ..., 0.01056047, 0.01133156,
        0.        ],
       [0.01632718, 1.        , 0.00719713, ..., 0.        , 0.        ,
        0.        ],
       [0.00999603, 0.00719713, 1.        , ..., 0.00635776, 0.        ,
        0.        ],
       ...,
       [0.01056047, 0.        , 0.00635776, ..., 1.        , 0.        ,
        0.        ],
       [0.01133156, 0.        , 0.        , ..., 0.        , 1.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        1.        ]])
```

We'll suggest items according to their similarity scores, However, our challenge arises from receiving these scores and subsequently needing to sort them. We rely on indices as reverse-map to associate the scores with their respective titles**.**

GETTING RECOMMENDATION

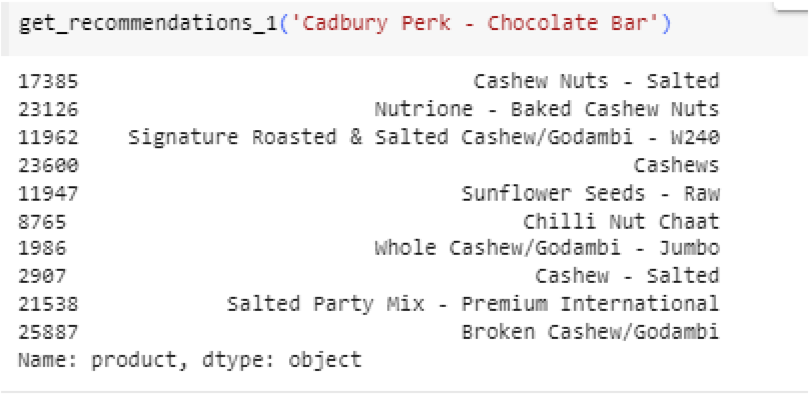
WATER BOTTLE:



```
get_recommendations_1('Water Bottle - Orange')
```

```
11320    Rectangular Plastic Container - With Lid, Mult...
11642                             Jar - With Lid, Yellow
26451     Round & Flat Storage Container - With lid, Green
6163     Premium Rectangular Plastic Container With Lid...
9546     Premium Round Plastic Container With Lid - Yellow
13959    Premium Rectangular Plastic Container With Lid...
19381    Premium Round & Flat Storage Container With Li...
24255      Premium Round Plastic Container With Lid - Blue
26067    Premium Round Plastic Container With Lid - Mul...
26074      Premium Round Plastic Container With Lid - Pink
Name: product, dtype: object
```

CADBURY PERK:

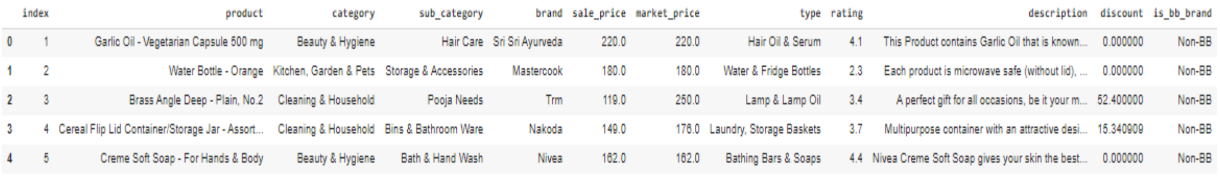


```
get_recommendations_1('Cadbury Perk - Chocolate Bar')
```

```
17385                               Cashew Nuts - Salted
23126                        Nutrione - Baked Cashew Nuts
11962    Signature Roasted & Salted Cashew/Godambi - W240
23600                                             Cashews
11947                               Sunflower Seeds - Raw
8765                                     Chilli Nut Chaat
1986                         Whole Cashew/Godambi - Jumbo
2907                                      Cashew - Salted
21538              Salted Party Mix - Premium International
25887                               Broken Cashew/Godambi
Name: product, dtype: object
```

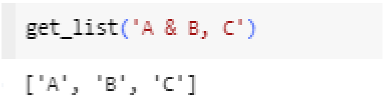
Here our search was Cadbury perk-chocolate, but we received recommendations for Cashew and Nuts as instead, so, we need to optimize this based on category, sub-category and brand

DATA COPY



	index	product	category	sub_category	brand	sale_price	market_price	type	rating	description	discount	is_bb_brand
0	1	Garlic Oil - Vegetarian Capsule 500 mg	Beauty & Hygiene	Hair Care	Sri Sri Ayurveda	220.0	220.0	Hair Oil & Serum	4.1	This Product contains Garlic Oil that is known...	0.000000	Non-BB
1	2	Water Bottle - Orange	Kitchen, Garden & Pets	Storage & Accessories	Mastercook	180.0	180.0	Water & Fridge Bottles	2.3	Each product is microwave safe (without lid), ...	0.000000	Non-BB
2	3	Brass Angle Deep - Plain, No.2	Cleaning & Household	Pooja Needs	Trm	119.0	250.0	Lamp & Lamp Oil	3.4	A perfect gift for all occasions, be it your m...	52.400000	Non-BB
3	4	Cereal Flip Lid Container/Storage Jar - Assort...	Cleaning & Household	Bins & Bathroom Ware	Nakoda	149.0	176.0	Laundry, Storage Baskets	3.7	Multipurpose container with an attractive desi...	15.340909	Non-BB
4	5	Creme Soft Soap - For Hands & Body	Beauty & Hygiene	Bath & Hand Wash	Nivea	162.0	162.0	Bathing Bars & Soaps	4.4	Nivea Creme Soft Soap gives your skin the best...	0.000000	Non-BB

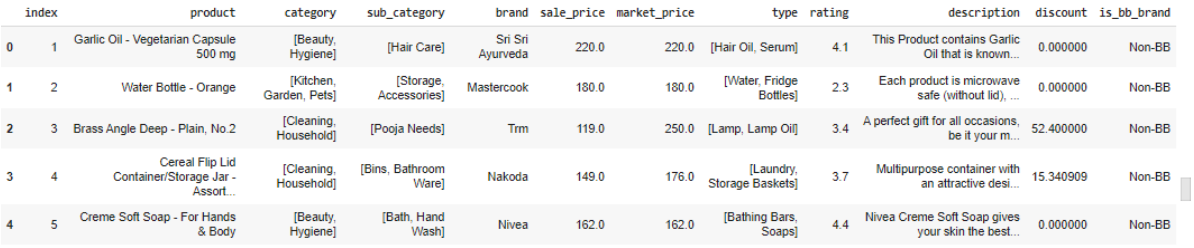
Take note that a single product may fall under several categories and subcategories, which are denoted by "&" symbols. We should parse these entries into a list to facilitate further processing.



```
get_list('A & B, C')
```

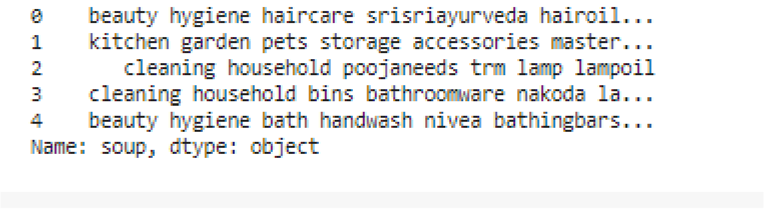
```
['A', 'B', 'C']
```

AFTER DATA COPY



	index	product	category	sub_category	brand	sale_price	market_price	type	rating	description	discount	is_bb_brand
0	1	Garlic Oil - Vegetarian Capsule 500 mg	[Beauty, Hygiene]	[Hair Care]	Sri Sri Ayurveda	220.0	220.0	[Hair Oil, Serum]	4.1	This Product contains Garlic Oil that is known...	0.000000	Non-BB
1	2	Water Bottle - Orange	[Kitchen, Garden, Pets]	[Storage, Accessories]	Mastercook	180.0	180.0	[Water, Fridge Bottles]	2.3	Each product is microwave safe (without lid), ...	0.000000	Non-BB
2	3	Brass Angle Deep - Plain, No.2	[Cleaning, Household]	[Pooja Needs]	Trm	119.0	250.0	[Lamp, Lamp Oil]	3.4	A perfect gift for all occasions, be it your m...	52.400000	Non-BB
3	4	Cereal Flip Lid Container/Storage Jar - Assort...	[Cleaning, Household]	[Bins, Bathroom Ware]	Nakoda	149.0	176.0	[Laundry, Storage Baskets]	3.7	Multipurpose container with an attractive desi...	15.340909	Non-BB
4	5	Creme Soft Soap - For Hands & Body	[Beauty, Hygiene]	[Bath, Hand Wash]	Nivea	162.0	162.0	[Bathing Bars, Soaps]	4.4	Nivea Creme Soft Soap gives your skin the best...	0.000000	Non-BB

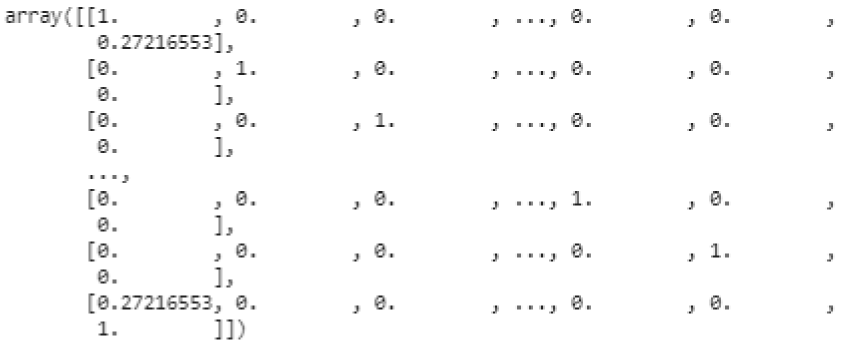
We will be converting all text into the lowercase and also eliminating the spaces between words for avoiding the duplicacy, This process ensure that our recommender doesn’t consider "Chocolate" of "CholocateIceCream" and "Chocolate Bar" are same.



```
0    beauty hygiene haircare srisriayurveda hairoil...
1    kitchen garden pets storage accessories master...
2       cleaning household poojaneeds trm lamp lampoil
3    cleaning household bins bathroomware nakoda la...
4    beauty hygiene bath handwash nivea bathingbars...
Name: soup, dtype: object
```

COUNT THE STRING VECTORS

We need to consider the count of string vectors and then compute the cosine similarity score.

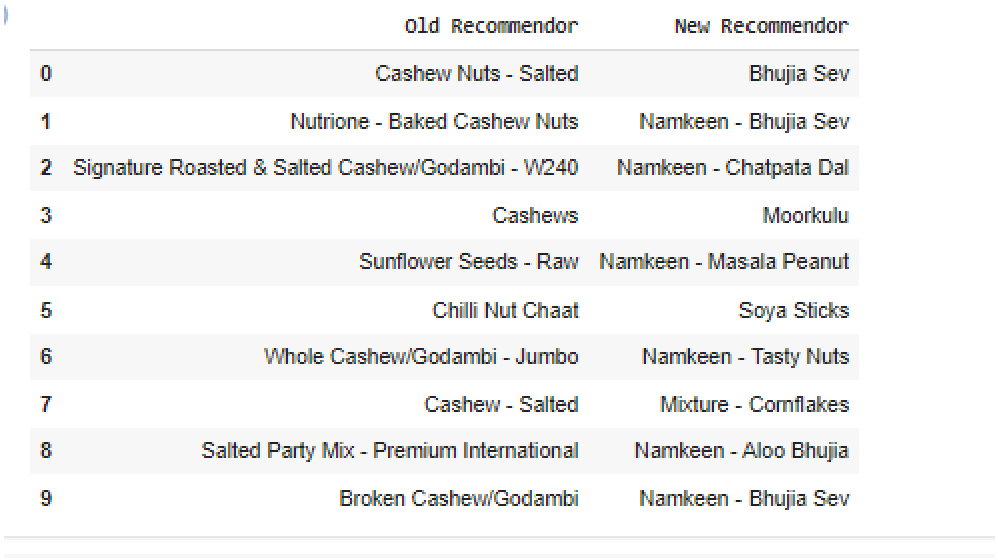


```
array([[1.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.27216553],
       [0.        , 1.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 1.        , ..., 0.        , 0.        ,
        0.        ],
       ...,
       [0.        , 0.        , 0.        , ..., 1.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 1.        ,
        0.        ],
       [0.27216553, 0.        , 0.        , ..., 0.        , 0.        ,
        1.        ]])
```

CONSINE SIMILARITY DOCUMENTATION:

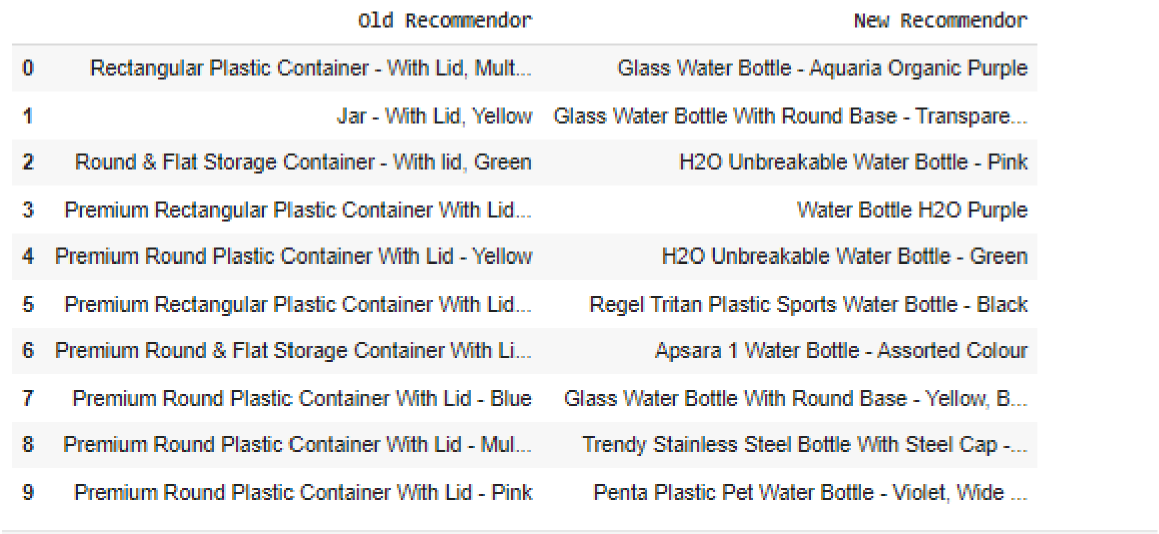
COMPARING OLD AND NEW RECOMMENDATION:

CADBURY PERK:



	Old Recommendor	New Recommendor
0	Cashew Nuts - Salted	Bhujia Sev
1	Nutrione - Baked Cashew Nuts	Namkeen - Bhujia Sev
2	Signature Roasted & Salted Cashew/Godambi - W240	Namkeen - Chatpata Dal
3	Cashews	Moorkulu
4	Sunflower Seeds - Raw	Namkeen - Masala Peanut
5	Chilli Nut Chaat	Soya Sticks
6	Whole Cashew/Godambi - Jumbo	Namkeen - Tasty Nuts
7	Cashew - Salted	Mixture - Cornflakes
8	Salted Party Mix - Premium International	Namkeen - Aloo Bhujia
9	Broken Cashew/Godambi	Namkeen - Bhujia Sev

WATER BOTTLE

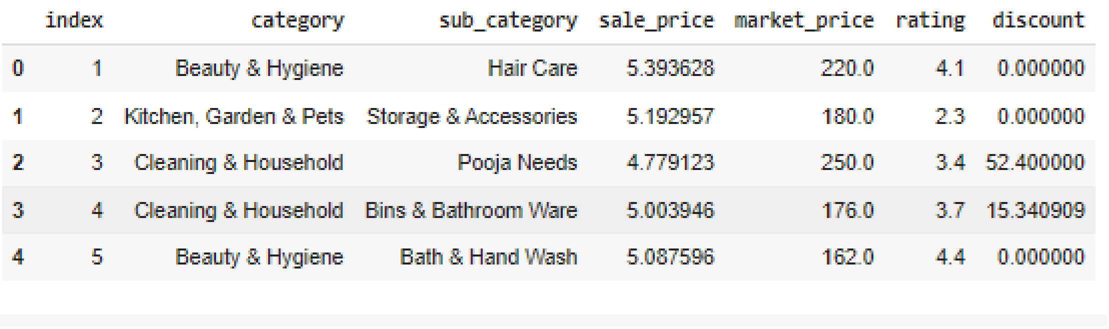


	Old Recommendor	New Recommendor
0	Rectangular Plastic Container - With Lid, Mult...	Glass Water Bottle - Aquaria Organic Purple
1	Jar - With Lid, Yellow	Glass Water Bottle With Round Base - Transpare...
2	Round & Flat Storage Container - With lid, Green	H2O Unbreakable Water Bottle - Pink
3	Premium Rectangular Plastic Container With Lid...	Water Bottle H2O Purple
4	Premium Round Plastic Container With Lid - Yellow	H2O Unbreakable Water Bottle - Green
5	Premium Rectangular Plastic Container With Lid...	Regel Tritan Plastic Sports Water Bottle - Black
6	Premium Round & Flat Storage Container With Li...	Apsara 1 Water Bottle - Assorted Colour
7	Premium Round Plastic Container With Lid - Blue	Glass Water Bottle With Round Base - Yellow, B...
8	Premium Round Plastic Container With Lid - Mul...	Trendy Stainless Steel Bottle With Steel Cap -...
9	Premium Round Plastic Container With Lid - Pink	Penta Plastic Pet Water Bottle - Violet, Wide ...

Figure 1.11

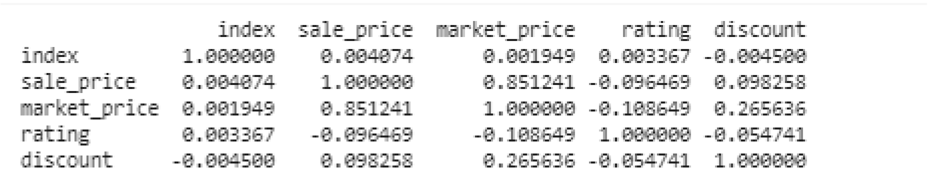
MODEL BUILDING:

READING DATASET



	index	category	sub_category	sale_price	market_price	rating	discount
0	1	Beauty & Hygiene	Hair Care	5.393628	220.0	4.1	0.000000
1	2	Kitchen, Garden & Pets	Storage & Accessories	5.192957	180.0	2.3	0.000000
2	3	Cleaning & Household	Pooja Needs	4.779123	250.0	3.4	52.400000
3	4	Cleaning & Household	Bins & Bathroom Ware	5.003946	176.0	3.7	15.340909
4	5	Beauty & Hygiene	Bath & Hand Wash	5.087596	162.0	4.4	0.000000

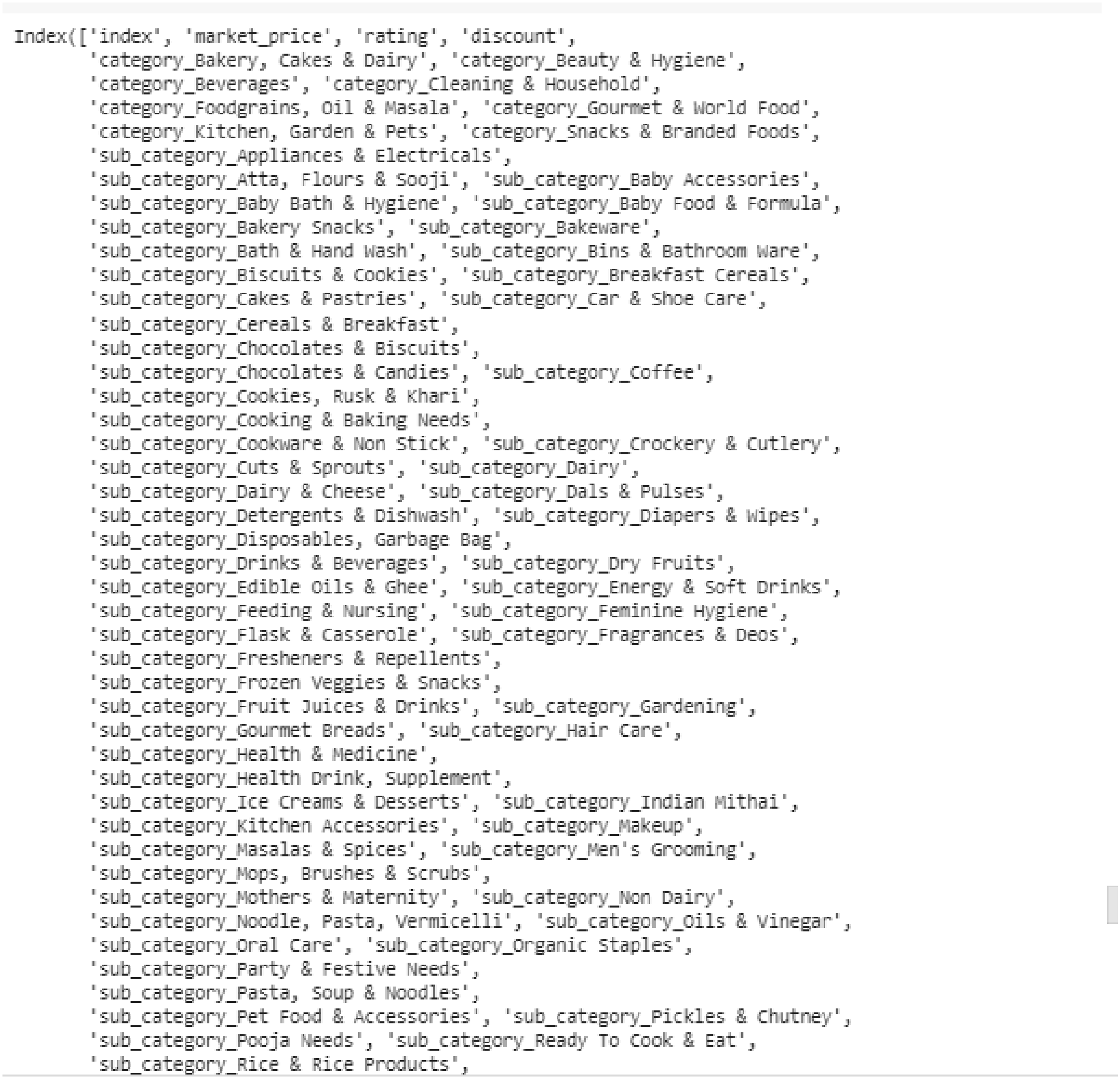
CORRELATION



```
                 index  sale_price  market_price    rating  discount
index         1.000000    0.004074      0.001949  0.003367 -0.004500
sale_price    0.004074    1.000000      0.851241 -0.096469  0.098258
market_price  0.001949    0.851241      1.000000 -0.108649  0.265636
rating        0.003367   -0.096469     -0.108649  1.000000 -0.054741
discount     -0.004500    0.098258      0.265636 -0.054741  1.000000
```

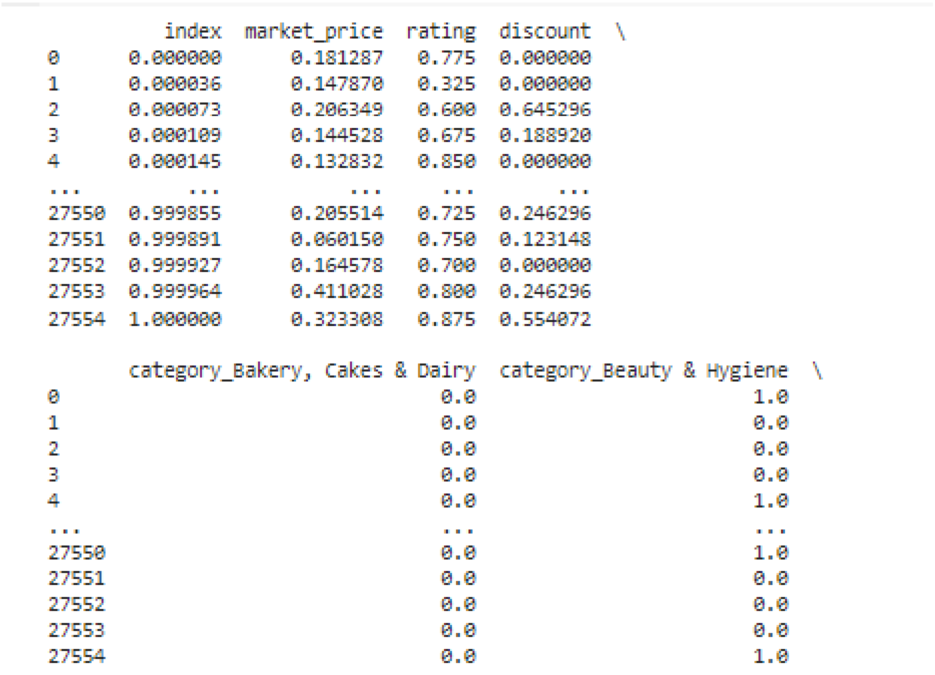
Figure 1.12

COULUMN



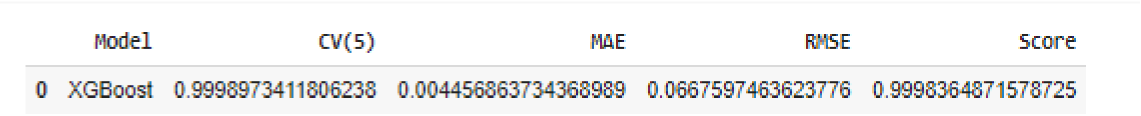
```
Index(['index', 'market_price', 'rating', 'discount',
       'category_Bakery, Cakes & Dairy', 'category_Beauty & Hygiene',
       'category_Beverages', 'category_Cleaning & Household',
       'category_Foodgrains, Oil & Masala', 'category_Gourmet & World Food',
       'category_Kitchen, Garden & Pets', 'category_Snacks & Branded Foods',
       'sub_category_Appliances & Electricals',
       'sub_category_Atta, Flours & Sooji', 'sub_category_Baby Accessories',
       'sub_category_Baby Bath & Hygiene', 'sub_category_Baby Food & Formula',
       'sub_category_Bakery Snacks', 'sub_category_Bakeware',
       'sub_category_Bath & Hand Wash', 'sub_category_Bins & Bathroom Ware',
       'sub_category_Biscuits & Cookies', 'sub_category_Breakfast Cereals',
       'sub_category_Cakes & Pastries', 'sub_category_Car & Shoe Care',
       'sub_category_Cereals & Breakfast',
       'sub_category_Chocolates & Biscuits',
       'sub_category_Chocolates & Candies', 'sub_category_Coffee',
       'sub_category_Cookies, Rusk & Khari',
       'sub_category_Cooking & Baking Needs',
       'sub_category_Cookware & Non Stick', 'sub_category_Crockery & Cutlery',
       'sub_category_Cuts & Sprouts', 'sub_category_Dairy',
       'sub_category_Dairy & Cheese', 'sub_category_Dals & Pulses',
       'sub_category_Detergents & Dishwash', 'sub_category_Diapers & Wipes',
       'sub_category_Disposables, Garbage Bag',
       'sub_category_Drinks & Beverages', 'sub_category_Dry Fruits',
       'sub_category_Edible Oils & Ghee', 'sub_category_Energy & Soft Drinks',
       'sub_category_Feeding & Nursing', 'sub_category_Feminine Hygiene',
       'sub_category_Flask & Casserole', 'sub_category_Fragrances & Deos',
       'sub_category_Fresheners & Repellents',
       'sub_category_Frozen Veggies & Snacks',
       'sub_category_Fruit Juices & Drinks', 'sub_category_Gardening',
       'sub_category_Gourmet Breads', 'sub_category_Hair Care',
       'sub_category_Health & Medicine',
       'sub_category_Health Drink, Supplement',
       'sub_category_Ice Creams & Desserts', 'sub_category_Indian Mithai',
       'sub_category_Kitchen Accessories', 'sub_category_Makeup',
       'sub_category_Masalas & Spices', 'sub_category_Men's Grooming',
       'sub_category_Mops, Brushes & Scrubs',
       'sub_category_Mothers & Maternity', 'sub_category_Non Dairy',
       'sub_category_Noodle, Pasta, Vermicelli', 'sub_category_Oils & Vinegar',
       'sub_category_Oral Care', 'sub_category_Organic Staples',
       'sub_category_Party & Festive Needs',
       'sub_category_Pasta, Soup & Noodles',
       'sub_category_Pet Food & Accessories', 'sub_category_Pickles & Chutney',
       'sub_category_Pooja Needs', 'sub_category_Ready To Cook & Eat',
       'sub_category_Rice & Rice Products',
```

SCALAR-FIT_TRANSFORM



```
          index  market_price  rating  discount  \
0      0.000000      0.181287   0.775  0.000000
1      0.000036      0.147870   0.325  0.000000
2      0.000073      0.206349   0.600  0.645296
3      0.000109      0.144528   0.675  0.188920
4      0.000145      0.132832   0.850  0.000000
...         ...           ...     ...       ...
27550  0.999855      0.205514   0.725  0.246296
27551  0.999891      0.060150   0.750  0.123148
27552  0.999927      0.164578   0.700  0.000000
27553  0.999964      0.411028   0.800  0.246296
27554  1.000000      0.323308   0.875  0.554072

       category_Bakery, Cakes & Dairy  category_Beauty & Hygiene  \
0                                 0.0                        1.0
1                                 0.0                        0.0
2                                 0.0                        0.0
3                                 0.0                        0.0
4                                 0.0                        1.0
...                               ...                        ...
27550                             0.0                        1.0
27551                             0.0                        0.0
27552                             0.0                        0.0
27553                             0.0                        0.0
27554                             0.0                        1.0
```

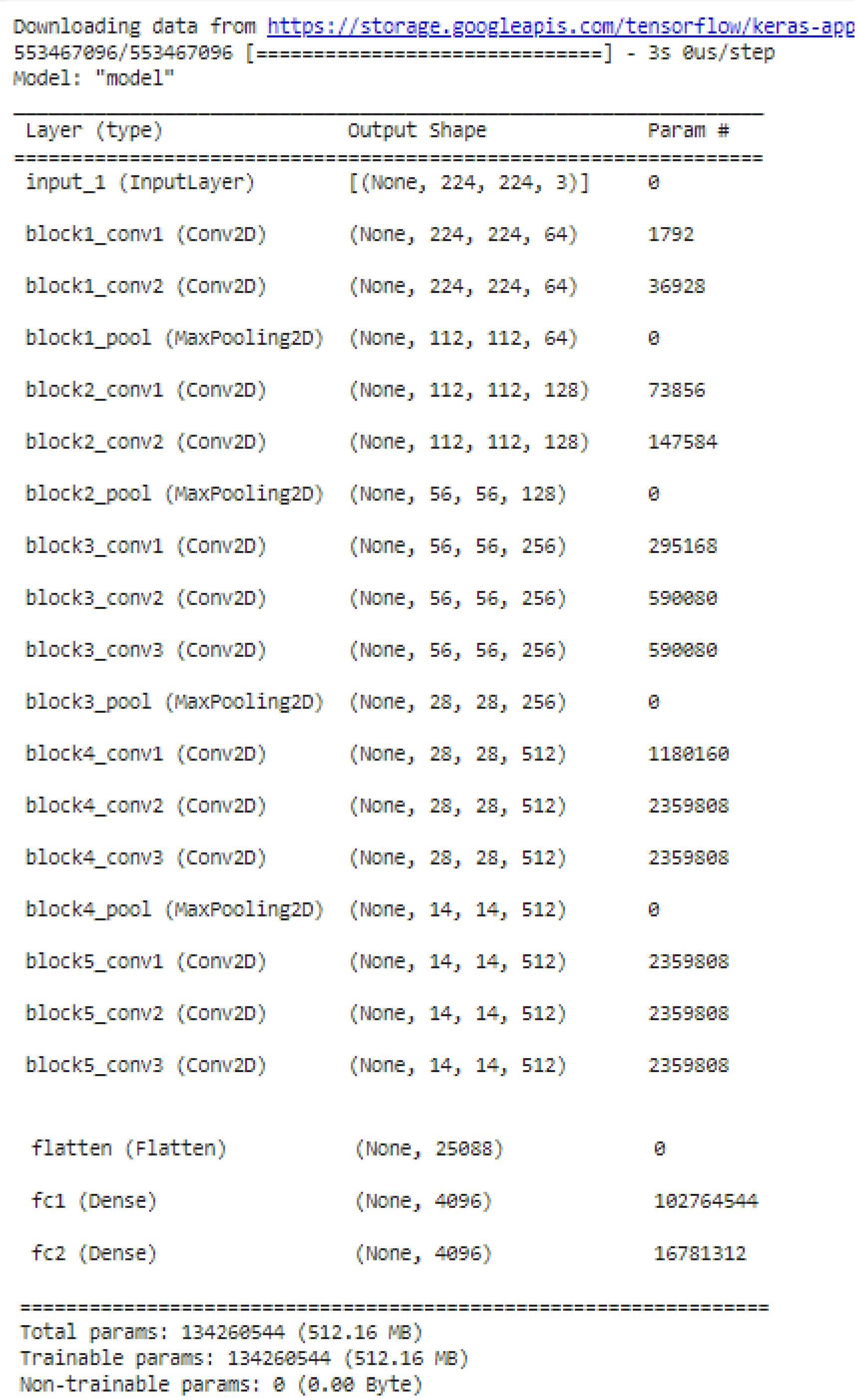
XG BOOSTER



	Model	CV(5)	MAE	RMSE	Score
0	XGBoost	0.9998973411806238	0.004456863734368989	0.0667597463623776	0.9998364871578725

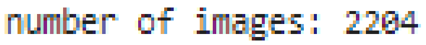
VISUALISING THE SIMILAR PRODUCT RECOMMENDATION:

LOAD THE VGG PRE –TRAINED MODEL FROM KERAS



```
Downloading data from https://storage.googleapis.com/tensorflow/keras-app
553467096/553467096 [==============================] - 3s 0us/step
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 224, 224, 3)]     0

 block1_conv1 (Conv2D)       (None, 224, 224, 64)      1792

 block1_conv2 (Conv2D)       (None, 224, 224, 64)      36928

 block1_pool (MaxPooling2D)  (None, 112, 112, 64)      0

 block2_conv1 (Conv2D)       (None, 112, 112, 128)     73856

 block2_conv2 (Conv2D)       (None, 112, 112, 128)     147584

 block2_pool (MaxPooling2D)  (None, 56, 56, 128)       0

 block3_conv1 (Conv2D)       (None, 56, 56, 256)       295168

 block3_conv2 (Conv2D)       (None, 56, 56, 256)       590080

 block3_conv3 (Conv2D)       (None, 56, 56, 256)       590080

 block3_pool (MaxPooling2D)  (None, 28, 28, 256)       0

 block4_conv1 (Conv2D)       (None, 28, 28, 512)       1180160

 block4_conv2 (Conv2D)       (None, 28, 28, 512)       2359808

 block4_conv3 (Conv2D)       (None, 28, 28, 512)       2359808

 block4_pool (MaxPooling2D)  (None, 14, 14, 512)       0

 block5_conv1 (Conv2D)       (None, 14, 14, 512)       2359808

 block5_conv2 (Conv2D)       (None, 14, 14, 512)       2359808

 block5_conv3 (Conv2D)       (None, 14, 14, 512)       2359808


 flatten (Flatten)           (None, 25088)             0

 fc1 (Dense)                 (None, 4096)              102764544

 fc2 (Dense)                 (None, 4096)              16781312

=================================================================
Total params: 134260544 (512.16 MB)
Trainable params: 134260544 (512.16 MB)
Non-trainable params: 0 (0.00 Byte)
```

GETTING THE IMAGE PATH:



```
number of images: 2204
```

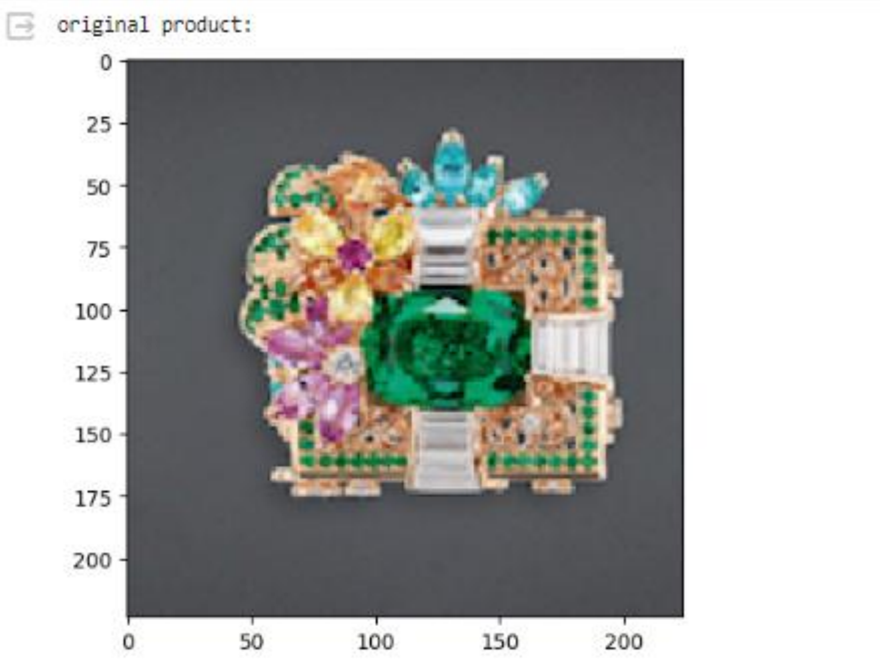
INPUT AN IMAGE INTO CNN:

Initially, we analyze the result obtained from inputting a single image into the CNN.

1. Loading the image.

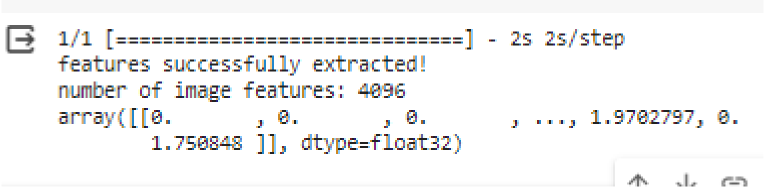
2. Preparing the image for CNN input.

3. Obtaining the CNN output that corresponds to the image's features.



```
original product:
```

GET THE EXTRACTED FEAUTURES:

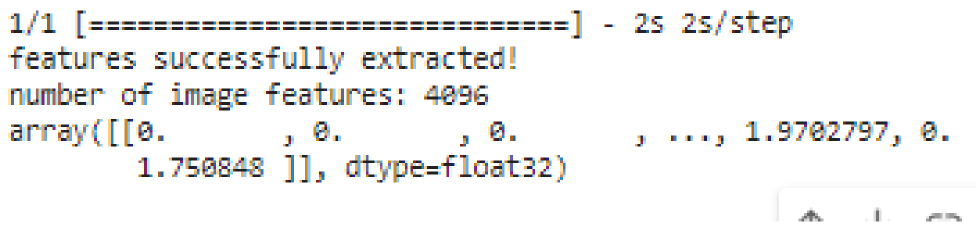


```
1/1 [==============================] - 2s 2s/step
features successfully extracted!
number of image features: 4096
array([[0.        , 0.        , 0.        , ..., 1.9702797, 0.        ,
        1.750848 ]], dtype=float32)
```

FEED ALL IMAGE IN TO THE CNN:

We successfully completed the feature extraction procedure for a single image. Next, let's do this process for all of our images!

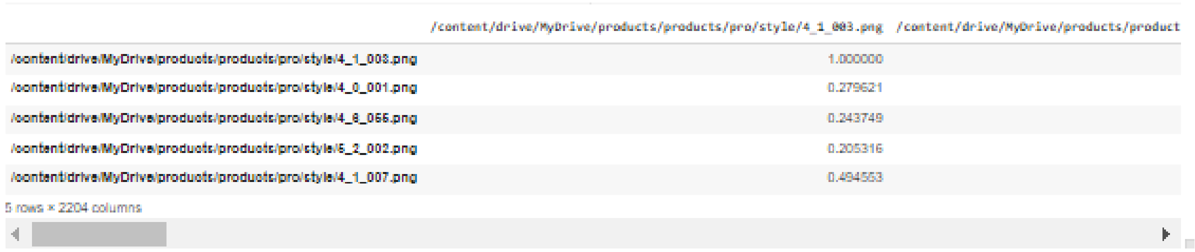
GET THE EXTRACTED IMAGE:



```
1/1 [==============================] - 2s 2s/step
features successfully extracted!
number of image features: 4096
array([[0.        , 0.        , 0.        , ..., 1.9702797, 0.
        1.750848 ]], dtype=float32)
```

COMPUTE COSINE SIMILARITIES:

Having obtained features for each image, we are now able to calculate similarity metrics for every pair of images, We will use her the cosine similarity metric in this context.

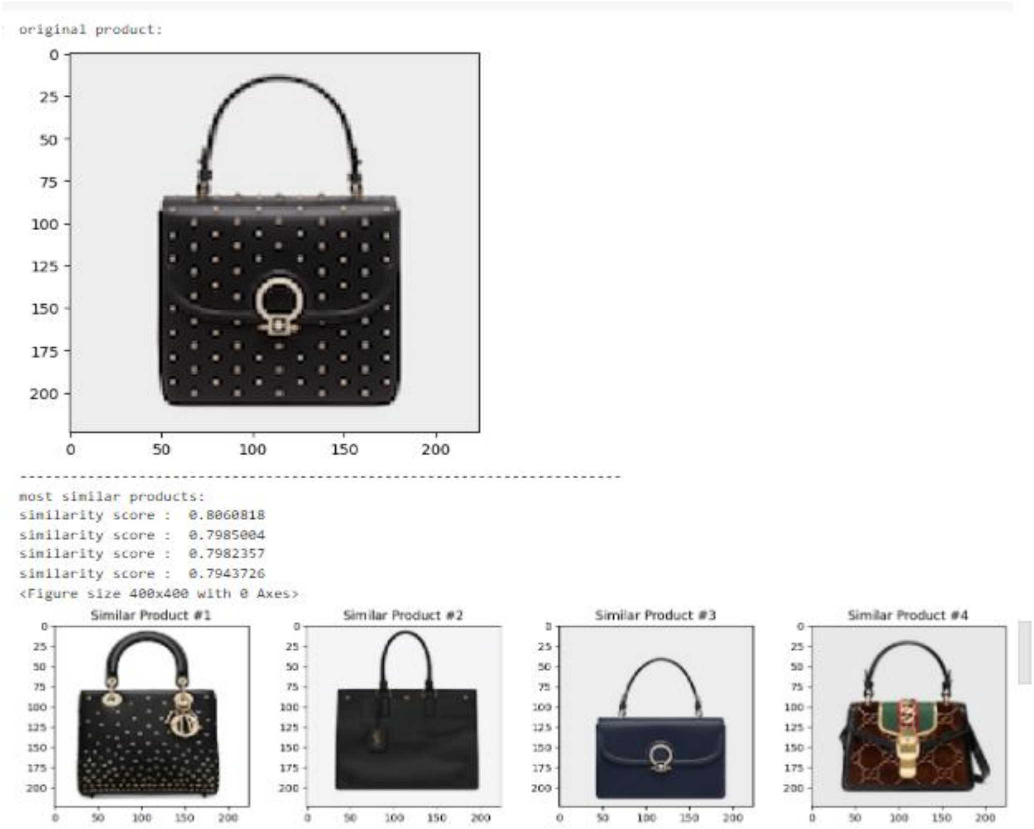


	/content/drive/MyDrive/products/products/pro/style/4_1_003.png	/content/drive/MyDrive/products/product
/content/drive/MyDrive/products/products/pro/style/4_1_003.png	1.000000	
/content/drive/MyDrive/products/products/pro/style/4_0_001.png	0.279621	
/content/drive/MyDrive/products/products/pro/style/4_8_066.png	0.243749	
/content/drive/MyDrive/products/products/pro/style/6_2_002.png	0.205316	
/content/drive/MyDrive/products/products/pro/style/4_1_007.png	0.494553	

5 rows × 2204 columns

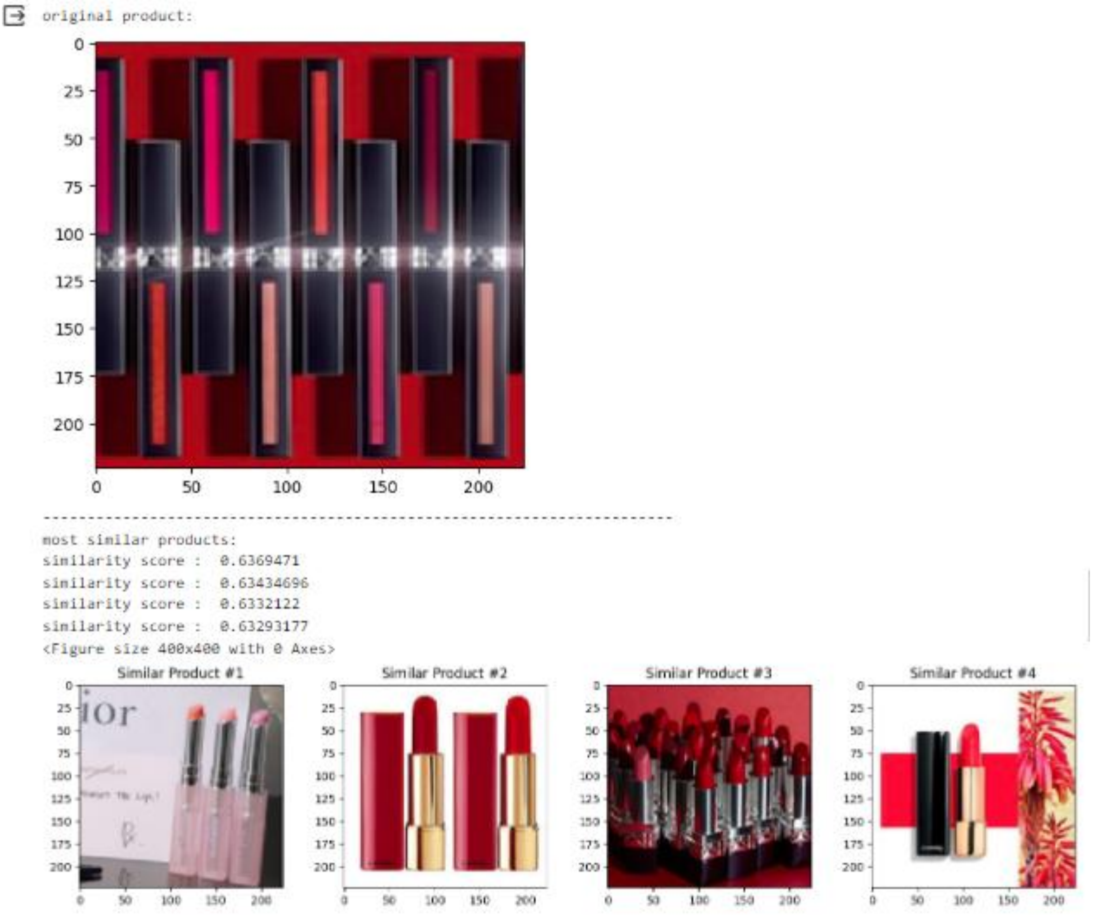
RETRIEVING MOST SIMILARITIES:

The last step involves creating a function that, given any product, identifies the most visually similar products.



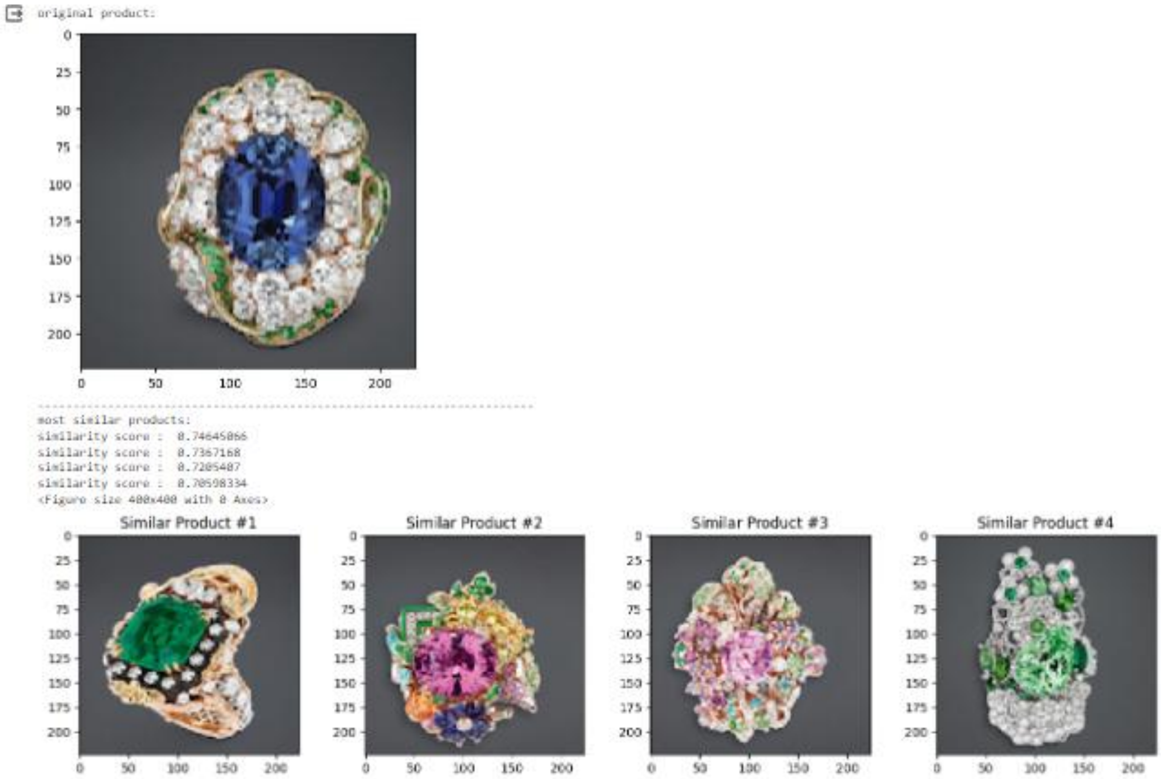
```
original product:
```

```
----------------------------------------------------------------------
most similar products:
similarity score :  0.8060818
similarity score :  0.7985004
similarity score :  0.7982357
similarity score :  0.7943726
<Figure size 400x400 with 0 Axes>
```

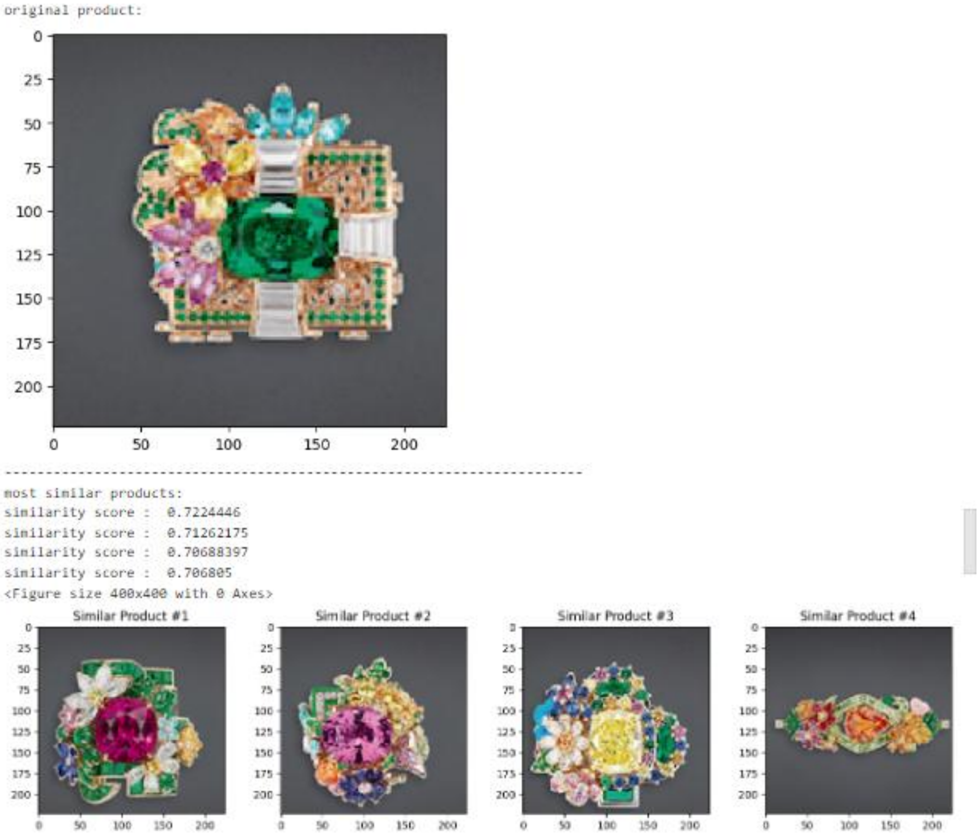


```
original product:
```

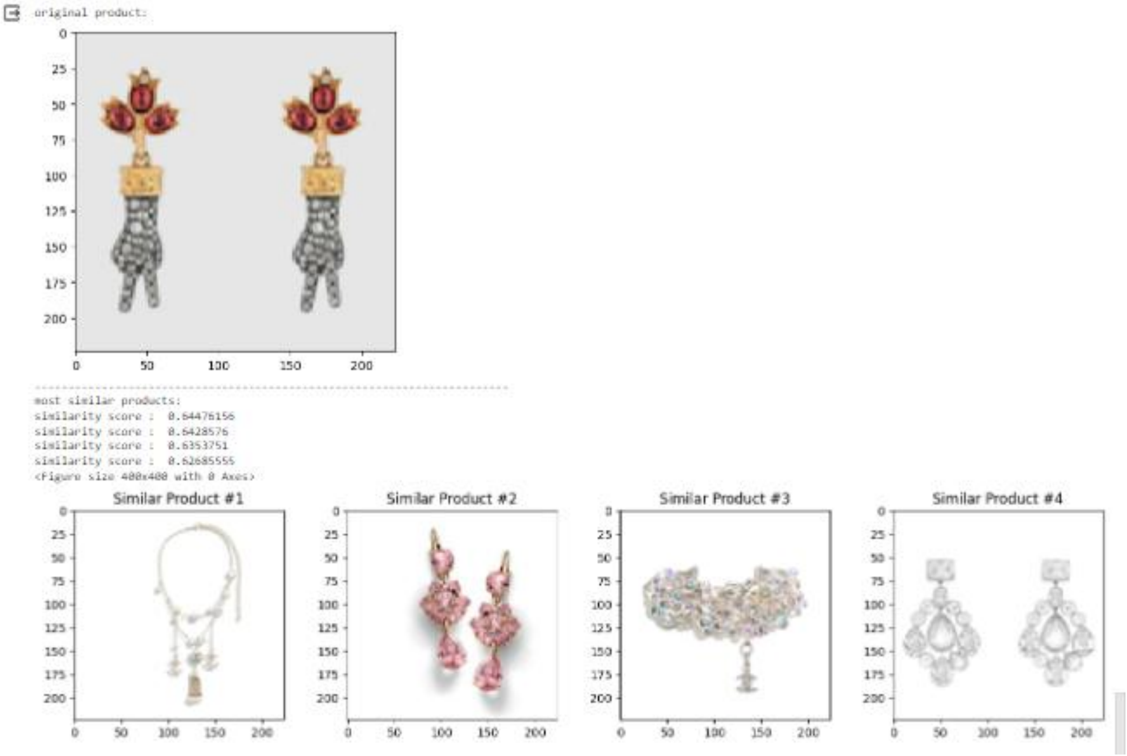
```
----------------------------------------------------------------------
most similar products:
similarity score :  0.6369471
similarity score :  0.63434696
similarity score :  0.6332122
similarity score :  0.63293177
<Figure size 400x400 with 0 Axes>
```



```
original product:
-------------------------------------------------------------------------
most similar products:
similarity score :  0.74645866
similarity score :  0.7367168
similarity score :  0.7205487
similarity score :  0.70598334
<Figure size 400x400 with 0 Axes>
```

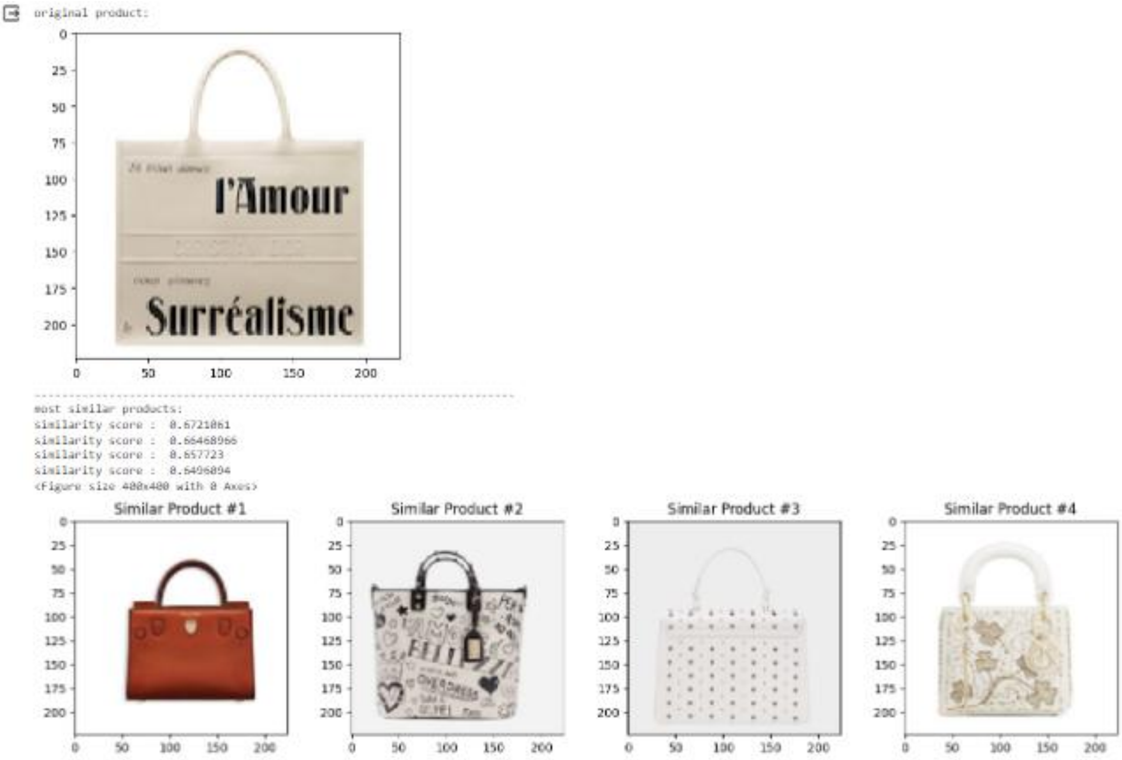


```
original product:
-------------------------------------------------------------------------
most similar products:
similarity score :  0.7224446
similarity score :  0.71262175
similarity score :  0.70688397
similarity score :  0.706805
<Figure size 400x400 with 0 Axes>
```



```
original product:
```

```
---------------------------------------------------------------------
most similar products:
similarity score :  0.64476156
similarity score :  0.6428576
similarity score :  0.6353751
similarity score :  0.62685555
<Figure size 400x400 with 0 Axes>
```



```
original product:
```

```
---------------------------------------------------------------------
most similar products:
similarity score :  0.6721061
similarity score :  0.66468966
similarity score :  0.657723
similarity score :  0.6496894
<Figure size 400x400 with 0 Axes>
```

Figure 1.13

7. ADVANTAGES AND DISADVANTAGES

ADVANTAGES OF PREDICTION

Enhanced Targeting: By analyzing the effectiveness of different platforms, messages, and offers in engaging and converting customers, businesses can strategically allocate their marketing resources toward specific customer segments.

Optimized ROI: Focusing on channels and campaigns with the potential for high return on investment enables businesses to maximize their marketing budgets through digital marketing predictions, potentially improving overall financial performance.

Competitive Edge: Staying ahead in the digital marketing landscape with predictive insights allows companies to surpass competitors, expand their customer base, and increase revenue by crafting more effective marketing strategies based on data-driven insights.

Revenue Growth: Leveraging similar product suggestions can drive sales growth as businesses can encourage additional purchases and boost revenue by recommending products aligned with customers' interests.

Enhanced Customer Experience: Providing personalized suggestions based on Improving the overall customer experience by tailoring recommendations to customer preferences promotes loyalty and repeat business, ensuring that the suggestions provided are aligned with customer requirements.

Time Efficiency: Similar product recommendations save customers time and effort by offering relevant suggestions without requiring additional search efforts, streamlining the shopping experience.

DISADVANTAGES OF PREDICTION

Inaccuracy of predictive model: The accuracy of predictive models, including those for online sales, cannot reach 100% due to various factors such as data quality and shifts in consumer behavior.

Limited Data : Depending on a business's size and scope, there may be limited data available for comprehensive predictive analysis.

Challenge in accounting for external factors: Predictive models struggle to incorporate external factors beyond a business's control, like unforeseen economic downturns impacting sales predictability.

Dependence on Predictive modeling: Over-reliance on predictive models can stifle creativity in sales strategies; they should be used alongside other analytical methods and approached with caution.

Ethical considerations: Ethical concerns arise regarding data privacy and bias in predictive model usage for sales, necessitating transparent and responsible data handling practices.

Limited Variety: Recommending only similar products can narrow customer exposure to product variety, potentially causing them to miss out on items they might have liked but weren't suggested.

Lack of personalization: In product recommendations may overlook individual customer preferences, resulting in less relevant suggestions.

8. CONCLUSION

In Conclusion, Machine Learning is profoundly impacting digital marketing predictions and similar product recommendations. Through the analysis of extensive client data, businesses leverage machine learning algorithms to forecast trends and offer tailored suggestions, enhancing their marketing strategies and bolstering profitability.

In the real world of digital marketing prediction, machine learning algorithms used various data points such as website activity, customer behavior, and purchasing history to forecast sales volumes and key performance indicators (KPIs). This enables enterprises to enhance their marketing approaches.

Similarly, in the context of similar product recommendations, machine learning algorithms analyze customer data to discern patterns and deliver personalized product suggestions. This enhances customer experiences, boosts satisfaction levels, and drives sales for businesses.

Overall, machine learning stands as a crucial tool for businesses seeking to optimize their digital marketing strategies and provide informed recommendations for their customers. By harnessing these technologies to analyze customer data, businesses can make data-driven decisions that lead to increased profits and sustained success.

REFERENCES

[1] Thomas Ragg, Wolfram Menzel, Walter Baum, Michael Wigbers, Bayesian learning for sales rate prediction for thousands of retailers, Neurocomputing, Volume 43, Issues 1– 4,2002,

[2] H. Yuan, W. Xu and M. Wang, "Can online user behavior improve the performance of sales prediction in E-commerce?," 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), San Diego, CA, USA, 2014, pp. 2347-2352.

[3] Malar, P. J. M. A. J. (2016). Innovative digital marketing trends 2016. 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT).

[4] Noaman M. Ali, Abdullah Alshahrani, Ahmed M. Alghamdi and Boris Novikov .Online Products Recommendations System Based on Analyzing Customers Reviews.

[5] T, G., Choudhary, R., & Prasad, S. (2018). Prediction of Sales Value in online shopping 63 using Linear Regression. 2018 4th International Conference on Computing Communication and Automation (ICCCA)

[6] Punam, K., Pamula, R., & Jain, P. K. (2018). A Two-Level Statistical Model for Big Mart Sales Prediction. 2018 International Conference on Computing, Power and Communication Technologies (GUCON).

[7] L. Huang, Z. Dou, Y. Hu and R. Huang, "Online Sales Prediction: An Analysis With Dependency SCOR-Topic Sentiment Model," in IEEE Access, vol. 7, pp. 79791-79797, 2019,

[8] Chen, T., Yin, H., Chen, H. et al. Online sales prediction via trend alignment-based multitask recurrent neural networks. KnowlInfSyst 62, 2139–2167 (2020).

[9] J Sekban - 2019.Applying machine learning algorithms in sales prediction.

[10] Michael Giering.Retail sales prediction and item recommendations using customer demographics at store level.

[11] Sharma, A. K., Goel, N., Rajput, J., & Bilal, M. (2020). An Intelligent Model for Predicting the Sales of a Product. 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)

[12] Mastanvali, K.Sankeerthana, B. Naveen, N.Vishal.PREDICTION OF ONLINE SALES USING LINEAR REGRESSION (IJCRT), Hydrabad, India.

[13] S. K. Punjabi, V. Shetty, S. Pranav and A. Yadav, "Sales Prediction using Online Sentiment with Regression Model," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 209-212

[14] P, R., & M, S. (2021). Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms. 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS).

[15] Niu, Y. (2020). Walmart Sales Forecasting using XGBoost algorithm and Feature engineering. 2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE).

[16] Singh, B., Kumar, P., Sharma, N., & Sharma, K. P. (2020). Sales Forecast for Amazon Sales with Time Series Modeling. 2020 First International Conference on Power, Control and Computing Technologies (ICPC2T).

[17] Purvika Bajaj, Renesa Ray, ShivaniShedge, ShravaniVidhate, Prof. Dr. 64 NikhilkumarShardoor. SALES PREDICTION USING MACHINE LEARNING ALGORITHMS (IRJET).

[18] P, R., & M, S. (2021). Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms. 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)

[19] Dr.K. Deepa, G.Raghuram. Sales Forecasting Using Machine Learning Models.

[20] Sreemathy, J; Prasath, N. In: 2022 Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) I-SMAC Library.