ENHANCING E-COMMERCE PERFORMANCE - A PYSPARK AND ML APPROACH

DHARUN PRUDHIV M
2033010

DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
M.Sc. (Decision and Computing sciences)
OF ANNA UNIVERSITY

November 2023

DEPARTMENT OF COMPUTING
COIMBATORE INSTITUTE OF TECHNOLOGY
(Autonomous Institution affiliated to Anna University)
COIMBATORE – 641014

COIMBATORE INSTITUTE OF TECHNOLOGY

(Autonomous Institution affiliated to Anna University)

COIMBATORE 641014

(Bonafide Certificate)

Project Work - I
Seventh Semester

ENHANCING E-COMMERCE PERFORMANCE - A PYSPARK AND ML APPROACH

Bonafide record of work done by
Dharun Prudhiv M
(Register No: 2033010)

Submitted in partial fulfilment of the requirements for the degree of
M.Sc. (Decision and Computing Sciences) of Anna University

November 2023

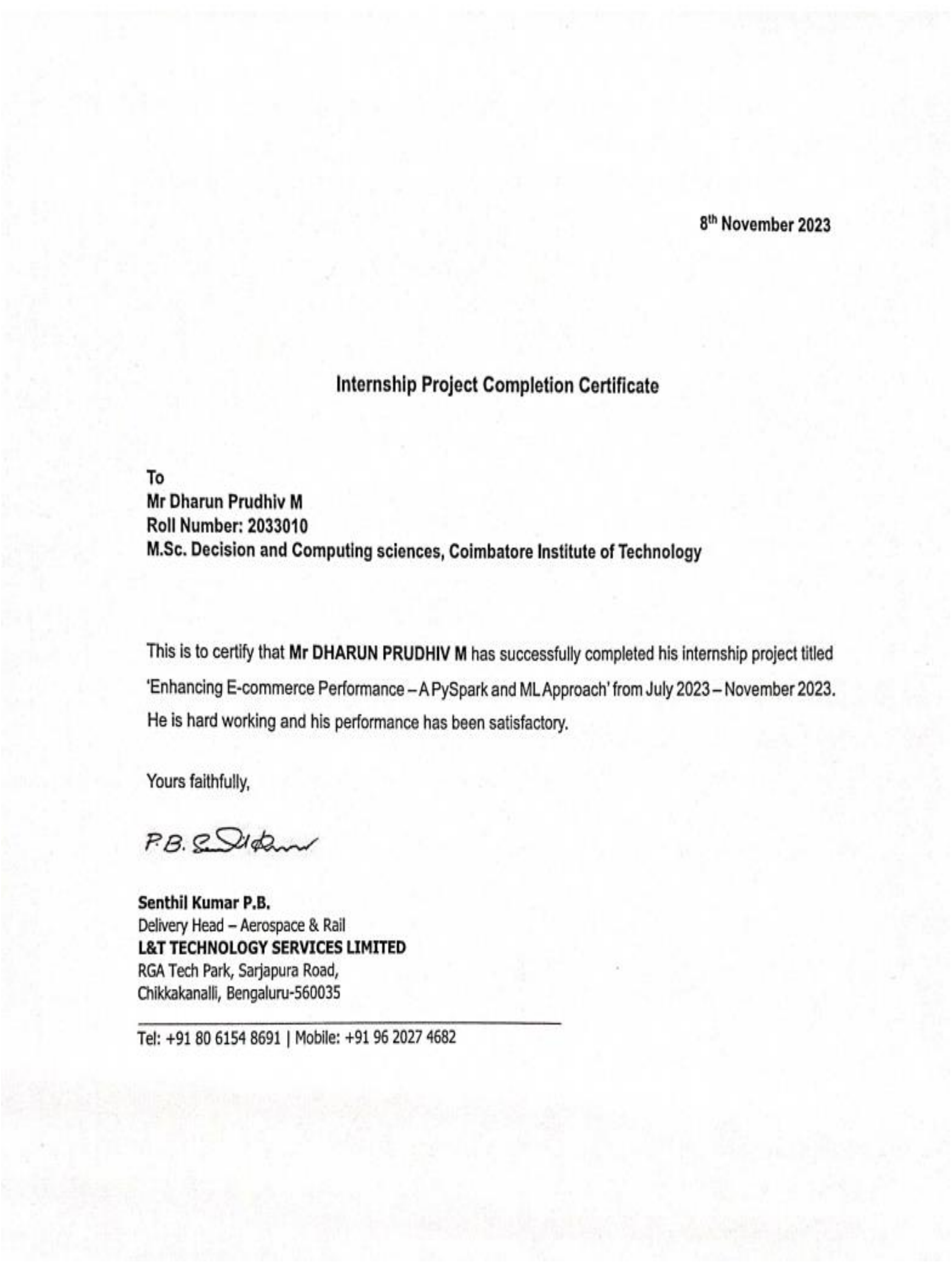
9/11/23
Faculty Guide

09/11/23
Head of the Department

Submitted for the viva-voce held on ________________

Internal Examiner

External Examiner



8th November 2023

Internship Project Completion Certificate

To
Mr Dharun Prudhiv M
Roll Number: 2033010
M.Sc. Decision and Computing sciences, Coimbatore Institute of Technology

This is to certify that **Mr DHARUN PRUDHIV M** has successfully completed his internship project titled 'Enhancing E-commerce Performance – A PySpark and ML Approach' from July 2023 – November 2023. He is hard working and his performance has been satisfactory.

Yours faithfully,

P.B.

Senthil Kumar P.B.
Delivery Head – Aerospace & Rail
L&T TECHNOLOGY SERVICES LIMITED
RGA Tech Park, Sarjapura Road,
Chikkakanalli, Bengaluru-560035

Tel: +91 80 6154 8691 | Mobile: +91 96 2027 4682

CONTENTS

ACKNOWLEDGEMENT

Apart from my efforts, the success of any project depends largely on the encouragement and guidelines of many others. I take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project.

I respect and thank **Dr. A. RAJESWARI, Principal**, Coimbatore Institute of Technology, for permitting me to undertake this project work at Crocus Technology Private Limited, Bangalore.

I express my sincere gratitude **Dr. K. SAKTHI MALA, Dean,** Department of Computing, Coimbatore Institute of Technology, Coimbatore, for her encouragement throughout this project.

I express my sincere gratitude to **Dr. A. KANNAMMAL, Professor and Head,** Department of Computing (Decision and Computing Sciences), Coimbatore Institute of Technology, Coimbatore, for her encouragement throughout this project.

I am indebted to my internal guide**, Dr. N. YAMUNA DEVI, Associate Professor**, Department of Computing (Decision and Computing Sciences), Coimbatore Institute of Technology, Coimbatore, for her constant support and guidance throughout the project work

I express my deep sense of gratitude to **Mr. P.B. SENTHIL KUMAR, Delivery Head**, L&T Technology Services, Bangalore for his invaluable guidance, support and suggestions throughout the course of this project work

I finally express at most gratitude to the almighty, my parents, my mentors and all of the team members for their help and support. They have been my motivators through my thick and thin.

SYNOPSIS

In the realm of modern e-commerce, data-driven decision-making has become indispensable for business success. This project, "Enhancing E-commerce Performance: A PySpark and ML Approach", has been meticulously crafted to harness the power of data to make informed choices in preparation for the upcoming Christmas season. The primary objective of this endeavor is to predict product profitability and promote those products with the highest potential for profit during the festive period.

The project encompasses a comprehensive data analysis pipeline, spanning exploratory data analysis (EDA), data preprocessing, PySpark-based analysis, and the development of both classification and regression models. These models serve a dual purpose: predicting product profitability or loss as well as estimating product prices or profits. By achieving these objectives, the project aims to provide valuable insights for informed marketing and product promotion strategies during the Christmas period.

The project encompasses several critical tasks, including data collection, preprocessing, feature engineering, and the implementation of multiple machine learning models. To improve predictive accuracy, ensemble techniques will be developed. Model evaluation will be conducted using essential metrics like AUC, precision, recall, and F1-score. Additionally, the project will focus on interpretability, analysing feature importance to provide valuable insights into the decision-making process, ultimately benefiting the client's understanding of the data.

Pandas, a versatile Python library, has played a pivotal role in data manipulation and analysis, streamlining tasks such as data cleaning and various data operations. Furthermore, Scikit-Learn, a renowned machine learning library, proved indispensable for implementing and experimenting with a range of machine learning algorithms, simplifying model training, evaluation, and feature engineering. PySpark, an open-source big data processing framework that integrates seamlessly with Python, has enabled the efficient handling and processing of large-scale datasets. Also developed and deployed machine learning models, enhancing the project's capabilities for data-driven decision-making and predictive insights.

PREFACE

CHAPTER I – INTRODUCTION gives an introduction of the organization for which the system was developed and also describes the objective and scope of the proposed system.

CHAPTER II - DATA MODELING AND EXPLORATION gives a detailed description about the data set used for analytics and the various models and techniques used, a comparison of those techniques and the proposed technique.

CHAPTER III - PREDICTIVE ANALYTICS PROCESS gives a detailed description about the process flow, the tools, Packages and libraries used for building the solution, and how the project is implemented.

CHAPTER IV - ANALYTICAL MODEL EVALUATION gives a detailed description about the performance measures used in the project.

CHAPTER V - ANALYSIS REPORTS AND INFERENCES gives a detailed description about reports and visual formats.

CHAPTER VI - CONCLUSION explains the reports and screens generated as part of the project.

CHAPTER I
INTRODUCTION

This section gives a detailed description of the organization for which the model is developed along with the explanation of the problem definition, goals and scope of the proposed model. This section also gives a descriptive summary of the data and specifies the methods, techniques and tools used in the development of the model and finally concludes with an inference.

1.1 ORGANIZATION PROFILE

L&T Technology Services (LTTS) is an Indian multinational technology company that provides engineering research and development (ER&D) services, headquartered in Vadodara. The company's business interests include automotive engineering, embedded system and semiconductor engineering, industrial internet of things, manufacturing plant engineering, and medical engineering.

1.2 PROBLEM STATEMENT

A prominent e-commerce platform is preparing for a digital marketing campaign designed to boost sales during the upcoming Christmas season. The primary objective of this campaign is to increase overall sales by a certain percentage. Additionally, the campaign aims to identify and highlight products with historically low sales volumes while emphasizing products that offer higher profit margins. To achieve this goal, an in-depth analysis of transactional data from past Christmas seasons will be undertaken to unveil trends, patterns, and customer preferences.

1.2.1 Objective

The primary objective of this project is to leverage data analysis techniques and machine learning models to optimize product promotion and pricing strategies for platform during the upcoming seasons. Specifically, the project aims to increase overall sales by a predefined percentage, identify and promote historically underperforming products, and emphasize products that yield higher profits. Through comprehensive data analysis, the project seeks to uncover trends, patterns, and customer preferences from past Christmas seasons, providing actionable insights to guide the digital marketing campaign and enhance sales performance.

1.2.2 scope

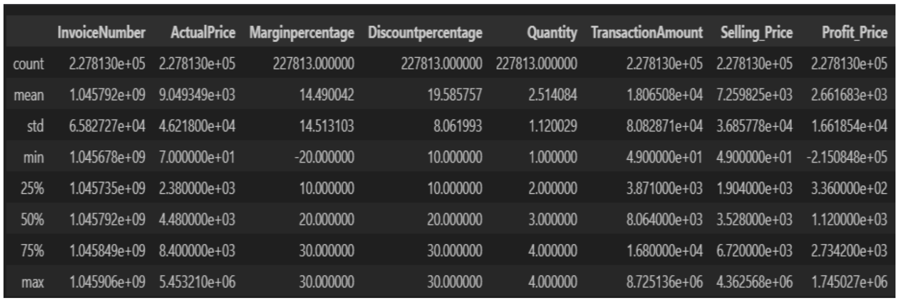
1. **Data Availability**: The project's effectiveness is heavily reliant on the availability and quality of historical transactional data. Inaccurate or incomplete data could impact the accuracy of insights and recommendations.
2. **Historical Data Relevance**: The assumption that past Christmas season data patterns will accurately predict future performance may not always hold true, as consumer behaviors and market dynamics can change over time.
3. **Model Accuracy**: The machine learning models' accuracy is contingent on the quality of the data and the choice of modeling techniques. Inaccurate models could lead to suboptimal product promotion and profitability predictions.
4. **Data Privacy and Security**: When dealing with transactional data, ensuring data privacy and security is crucial. Limitations may arise if there are constraints on accessing or sharing certain data due to privacy regulations or security concerns.
5. **Model Interpretability**: Complex machine learning models may lack interpretability, making it challenging to understand the rationale behind certain predictions and recommendations.

1.3 DESCRIPTIVE STATSTICAL SUMMARY

In this section, a detailed technical overview is presented for the descriptive statistics extracted from the historical transactional dataset. Investigation will delve into sales trends, product performance metrics, customer behaviour insights, time-based analysis, and correlations between variables. This comprehensive examination equips with the technical insights required to make data-driven decisions for the upcoming seasons sales campaign, ensuring precise and optimized strategies for maximizing sales and profitability. Table 1.1 represents a summary of key statistics for several variables in the dataset

- The average actual price of products is approximately 9,049 units, while the average selling price is around 7,260 units.
- There is a notable difference between the average actual price and the average selling price, indicating that discounts or promotional pricing strategies may be commonly employed.
- The average quantity of products sold per transaction is around 2.51 units.
- The minimum quantity is 1 unit, while the maximum is 4 units, with most transactions falling within this range.
- The minimum profit price is positive, suggesting that, on average, products are sold at a profit.
- The average margin percentage is approximately 14.49%, suggesting a reasonable profit margin on average.
- The average discount percentage is approximately 19.59%, indicating that discounts are frequently applied to products.

Table 1.1 Descriptive Statistics



	InvoiceNumber	ActualPrice	Marginpercentage	Discountpercentage	Quantity	TransactionAmount	Selling_Price	Profit_Price
count	2.278130e+05	2.278130e+05	227813.000000	227813.000000	227813.000000	2.278130e+05	2.278130e+05	2.278130e+05
mean	1.045792e+09	9.049349e+03	14.490042	19.585757	2.514084	1.806508e+04	7.259825e+03	2.661683e+03
std	6.582727e+04	4.621800e+04	14.513103	8.061993	1.120029	8.082871e+04	3.685778e+04	1.661854e+04
min	1.045678e+09	7.000000e+01	-20.000000	10.000000	1.000000	4.900000e+01	4.900000e+01	-2.150848e+05
25%	1.045735e+09	2.380000e+03	10.000000	10.000000	2.000000	3.871000e+03	1.904000e+03	3.360000e+02
50%	1.045792e+09	4.480000e+03	20.000000	20.000000	3.000000	8.064000e+03	3.528000e+03	1.120000e+03
75%	1.045849e+09	8.400000e+03	30.000000	30.000000	4.000000	1.680000e+04	6.720000e+03	2.734200e+03
max	1.045906e+09	5.453210e+06	30.000000	30.000000	4.000000	8.725136e+06	4.362568e+06	1.745027e+06

1.4 OVERVIEW OF PREDICTIVE ANALYSIS

1.4.1 Methodology

The project adopts a comprehensive methodology aligned with real-world e-commerce practices to optimize product promotion and pricing strategies. It begins with extensive data collection, encompassing transaction records, product details, customer demographics, and historical sales data, forming the foundational dataset. Subsequently, rigorous data preprocessing and cleaning are applied to ensure data consistency and quality. Exploratory Data Analysis (EDA) is performed using, Python and PySpark analysis is done further to uncover hidden trends and patterns within the data, facilitating a holistic understanding of its characteristics. Feature engineering is employed to enhance model performance by creating and modifying features to capture relevant information. Multiple machine learning models are then implemented to predict profitability, product prices classification models for profitability prediction and regression models for price estimation. Ensemble techniques are leveraged to improve predictive accuracy and model robustness. Model evaluation is carried out using key metrics like AUC, precision, recall, and F1-score to ensure alignment with project objectives. Feature importance analysis is conducted to provide valuable insights into product promotion and pricing strategies, aiding in understanding the factors influencing customer choices.

1.4.2 Tools Used

1. Python Programming
2. PySpark
3. scikit-learn
4. Pandas
5. Matplotlib, Seaborn, Squarify
6. Multivariate Ensembles
7. Data Preprocessing Tools
8. VS code – Jupyter Notebook

1.5 INFERENCES SUMMARY

In summary, the project "Enhancing E-commerce Performance: A PySpark and ML Approach" employs a range of machine learning models, including multiple linear regressor and decision tree regressor for regression tasks, as well as XGBoost, KNN classifier, Gradient Boosting, and RandomForest for classification purposes. The classification models exhibit strong performance, with the Random Forest Classifier and KNN Classifier achieving an accuracy of 98% and high precision, recall, and F1 scores. Similarly, the regression models, including Linear Regression and Decision Tree Regression, demonstrate their effectiveness with low mean squared errors (MSE) and high R2 scores. These results underline the project's success in providing valuable insights and predictive capabilities for optimizing e-commerce strategies and pricing decisions.

CHAPTER II

DATA MODELING AND EXPLORATION

2.1 PROBLEM ANALYSIS

2.1.1 Problem Understanding

The primary challenge in this project, "Enhancing E-commerce Performance: A PySpark and ML Approach" lies in optimizing product promotion and pricing strategies for customer during the upcoming seasons. The overarching goal is to enhance overall sales performance by a defined percentage while also targeting products with historically low sales and prioritizing those with higher profit margins. To achieve these objectives, deciphering intricate patterns within transactional data, customer preferences, and product performance is imperative.

2.1.2 Business Understanding

At the core of this project, titled 'Enhancing E-commerce Performance - A PySpark and ML Approach,' lies a profound recognition of the pivotal role that data-driven decision-making plays in the fiercely competitive e-commerce arena. In the realm of online retail, where consumer preferences and market dynamics continually shift, the ability to make informed decisions is the cornerstone of success. This section highlights the critical significance of comprehending the e-commerce landscape and the far-reaching implications of the data analysis endeavors.

Furthermore, the insights garnered from this project extend beyond the confines of e-commerce. The principles of market research, analysis of consumer behavior, and data-driven decision-making hold immense importance not only for e-commerce but also for industries where understanding customer behavior and preferences stands as a paramount consideration.

2.1.3 Feature Identification

In the realm of e-commerce data analysis, the process of feature identification is a critical step that forms the bedrock of the project's success. A comprehensive set of features tailored to capture the intricate dynamics,operations, particularly during the upcoming Christmas sales campaign.

This set of features includes "Profit_Price" and "Selling_Price," which serve as key metrics for assessing pricing strategies and profit margins. These features provide insights into the estimated profit generated per product sold and the actual selling price of products, enabling a deep dive into pricing dynamics.

Temporal aspects are also considered with features like "Day of Week," "Month," and "Year," instrumental in uncovering temporal patterns in customer behavior and sales trends. Help us discern seasonality and understand the impact of timing on sales performance, aiding in the strategic planning of promotions and campaigns.

Additionally, the binary "Profit/Loss" attribute categorizes transactions as either resulting in profit (1) or loss (0). This feature offers a clear view of the overall profitability of the e-commerce store during the Christmas season, serving as a fundamental metric for assessing the campaign's success.

Existing features such as product details, customer demographics, and transactional information are retained, ensuring a holistic view of customer behavior. These meticulously identified features collectively form a powerful toolkit, enabling us to extract actionable insights from the data and guide the client in data-driven decision-making to enhance sales and profitability during the festive season.

2.2 DATA MODEL

2.2.1 Data Collection

In the project, " Enhancing E-commerce Performance: A PySpark and ML Approach”, data collection plays a pivotal role in gaining insights into customer behavior and optimizing client’s strategies for the approaching Christmas season. The dataset encompasses a diverse range of attributes, each serving a distinct purpose in revealing the dynamics of e-commerce during this festive period.

Firstly, transactional data encompassing details such as "InvoiceNumber," "ProductCode," and "CustomerId." These attributes form the backbone of analysis, enabling us to track individual purchases, products, and customer identifiers, which is fundamental for understanding sales patterns.

Secondly, product details like "Brand," "Category," and "SubCategory" provide essential insights into the product hierarchy and categorization. This information is invaluable for discerning product preferences and effectively managing inventory.

Pricing and profitability metrics are represented by attributes such as "ActualPrice," "Selling_Price," and "Profit_Price." These columns allow us to assess pricing strategies and profitability by gauging actual and estimated profits per product, supporting informed pricing decisions.

2.2.2 Data Preparation

Prior to initiating the analysis and modeling phases of the project, systematic dataset preparation is imperative to ensure it is well-prepared for generating meaningful insights. Data preparation involves a series of crucial steps aimed at cleaning, preprocessing, and structuring the data for optimal utilization.

The first step in this process is data cleaning, involving meticulous identification and rectification of any inconsistencies or errors within the dataset. This encompasses addressing missing values, handling duplicates, and ensuring data quality is of the utmost standard.

Following data cleaning, feature engineering takes center stage. This involves crafting new attributes or modifying existing ones to extract more meaningful information from the dataset. Categorical variables like "Brand," "Category," and "Country" are converted into numerical form to make them compatible with machine learning models. Techniques such as label encoding are employed to represent these variables numerically while retaining their relevance.

Normalization is employed on numerical attributes, particularly those with differing scales. This procedure standardizes these attributes to a common range, preventing potential bias in the models. Standardizing these numerical features ensures that each attribute contributes proportionally to analysis, regardless of its initial scale.

Additionally, the dataset is partitioned into training and testing sets to enable accurate evaluation of model performance. This division enables model validation on unseen data and guards against overfitting.

2.3 EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis is an approach to analyses the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representation. It is crucial to understand it in depth before the perform data analysis and run the data through an algorithm. It is need to know the patterns in data and determine which variables are important and which do not play a significant role in the output. Further, some variables may have correlations with other variables. It also needs to recognize errors in the data.

All of this can be done with Exploratory Data Analysis. It helps to gather insights and make better sense of the data, and removes irregularities and unnecessary values from data.

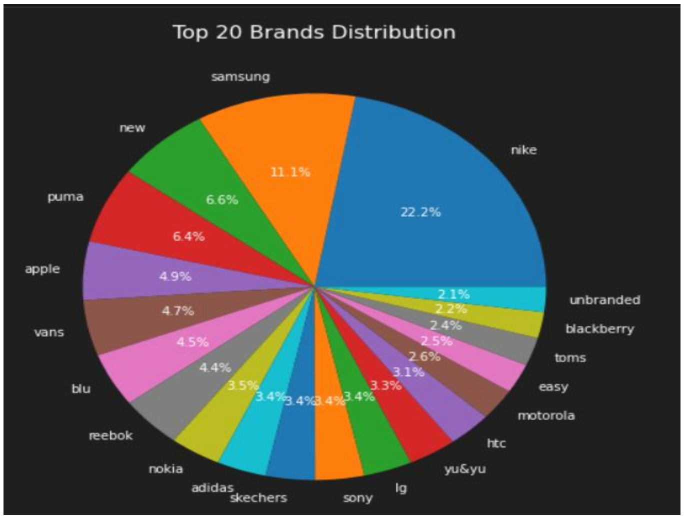


Figure 2.1 Pie-Chart

Figure 2.1 represents the distribution of Top 20 brands

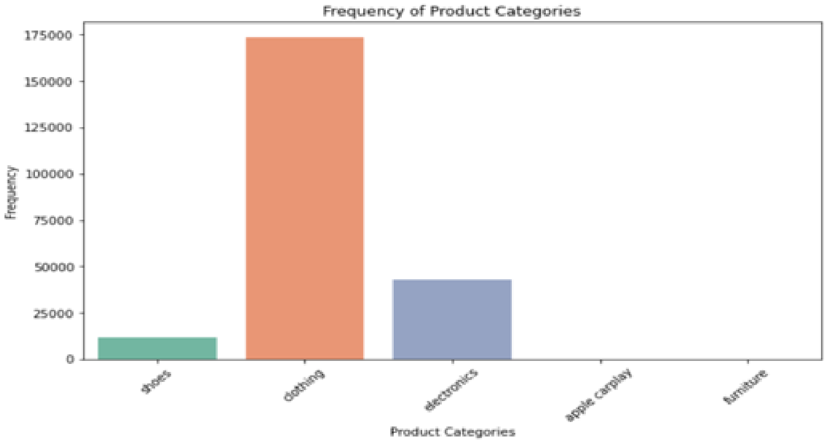


Figure 2.2 Bar-chart

Figure 2.2 represents the frequency of product categories

2.3.1 Inferences

- The dataset encompasses products from 3,578 distinct brands, reflecting the diversity of brands within the store's inventory.
- The "Clothing" category has the highest count of products, followed by "Electronics" and "Shoes," indicating that these are the most prominent product categories in the store.
- France has the highest transaction count, with 25,862 transactions, followed by Belgium (14,319), Israel (14,129), while Brazil and Iceland have the lowest transaction counts, each with 11,138 transactions. The dataset includes transactions from a total of 17 countries.
- In terms of transaction amounts, "Clothing" dominates the dataset, accounting for approximately 76.1% of the total transaction amount, followed by "Electronics" at 18.8%, and "Shoes" at 5.1%.
- Samsung, Nike, Blackberry, Toms, Motorola, HTC are among the brands with the highest percentage of transaction counts in the dataset, indicating their popularity among customers.
- Thursdays and fridays experience higher transaction counts compared to other days of the week, suggesting that these days are particularly busy for the store.
- The top brands in terms of transaction count include Nike, Samsung, Puma, Apple, Vans, Adidas, Blackberry, Sony, and Blu, indicating that these brands are among the most preferred by customers.

2.3.2 Inferences – Correlation Analysis

- ActualPrice and TransactionAmount: Strong positive correlation. Higher actual prices are associated with higher transaction amounts.
- Marginpercentage and Profit_Price: Moderate positive correlation. Higher margin percentages tend to result in higher profit prices.
- Marginpercentage and Discountpercentage: Weak negative correlation. Products with higher margin percentages receive lower discounts.

- Quantity and TransactionAmount: Positive correlation. Larger quantities sold in a transaction are linked to higher transaction amounts.
- Profit/Loss, Marginpercentage, and Selling_Price: Strong positive correlation. Higher profit margins and selling prices lead to higher profits or reduced losses.
- The representation of the strength and direction of relationships between variables is succinctly captured through correlation analysis, signifying the strength and direction of a linear relationship between two continuous variables is represented in figure2.3.

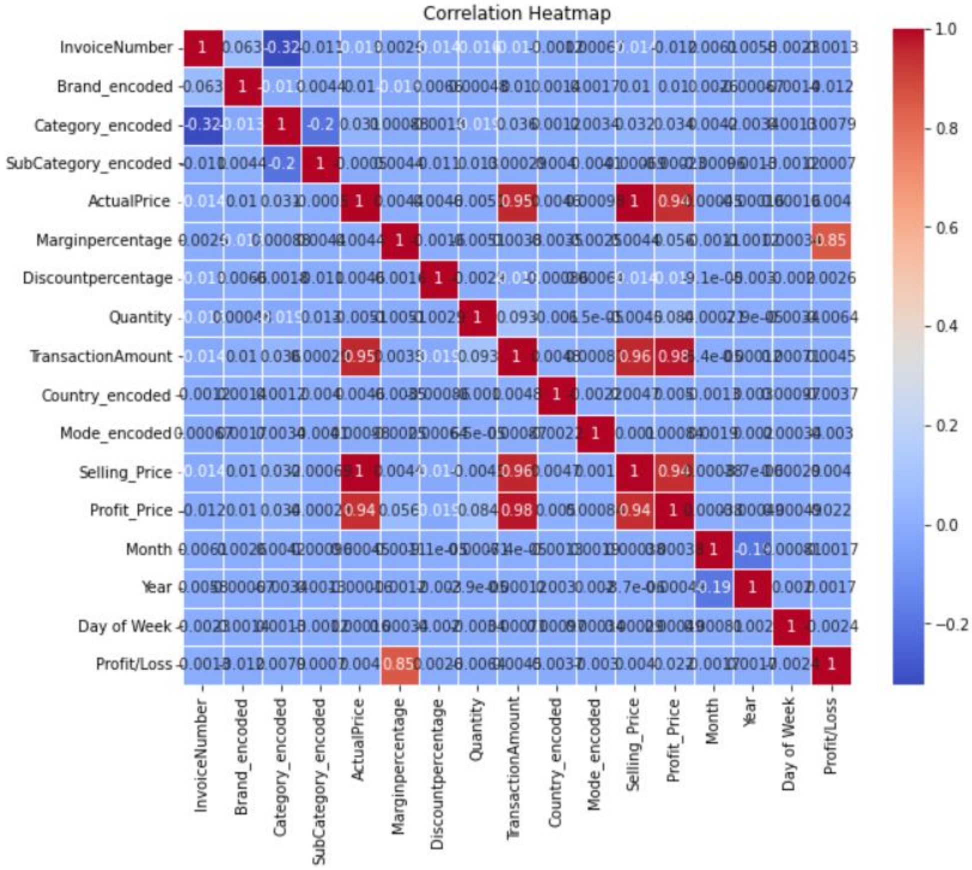


Figure 2.3 Correlation Heatmap

2.3.3 Inferences – Temporal Analysis

- **Monthly Transaction Amount**
    - March consistently generates the highest monthly transaction amount.
    - January follows closely behind, indicating robust sales revenue during these months.
- **Highest Monthly Product Sales**
    - March records the highest number of products sold, with 65,314 units.
    - January and February closely trail with sales counts of 56,673 and 56,824, respectively.
- **New Customer Acquisition**
    - There is a noticeable concentration of new customer acquisition at the beginning of each year, suggesting a surge in new customers during this period.
- **Quarterly Transaction Amount**
    - Across all years, the 1st quarter (Q1) consistently exhibits the highest total transaction amount.
- **Yearly Sales Trends**
    - In 2020, the 4th quarter (Q4) sees a significant upswing in sales.
    - In 2021, the 1st quarter (Q1) records the highest sales performance.
    - In 2022, the 2nd quarter (Q2) stands out with the highest sales figures.
    - Sales generally exhibit growth after the 3rd quarter (Q3) in each year.

These insights shed light on the seasonality and trends in sales, helping to identify peak months for revenue generation and recurring patterns in customer acquisition and sales performance.

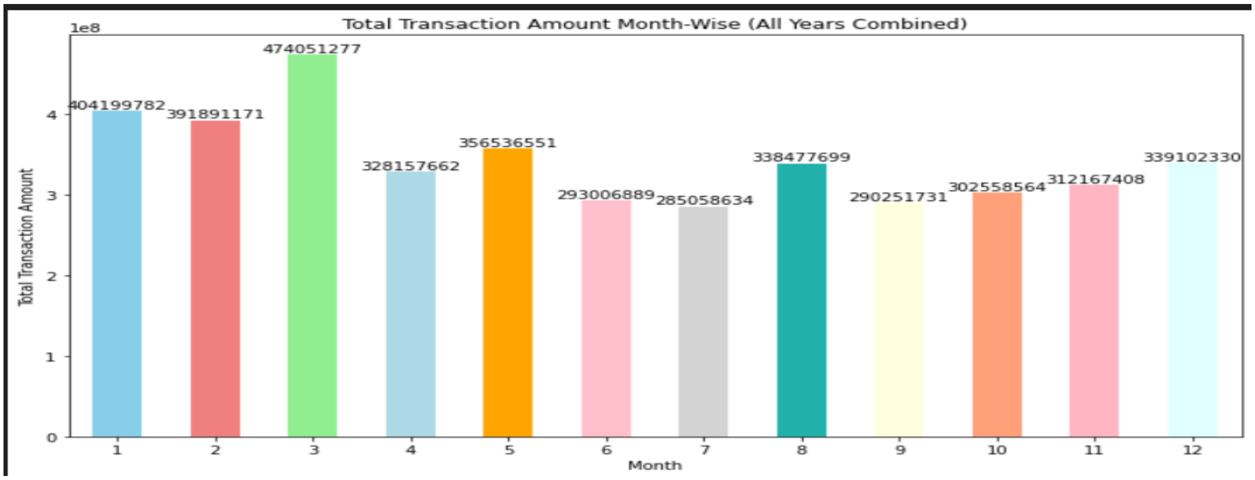


Figure 2.4 Month-wise Total Transaction Amount

Figure 2.4 represents the total transaction amount month-wise.

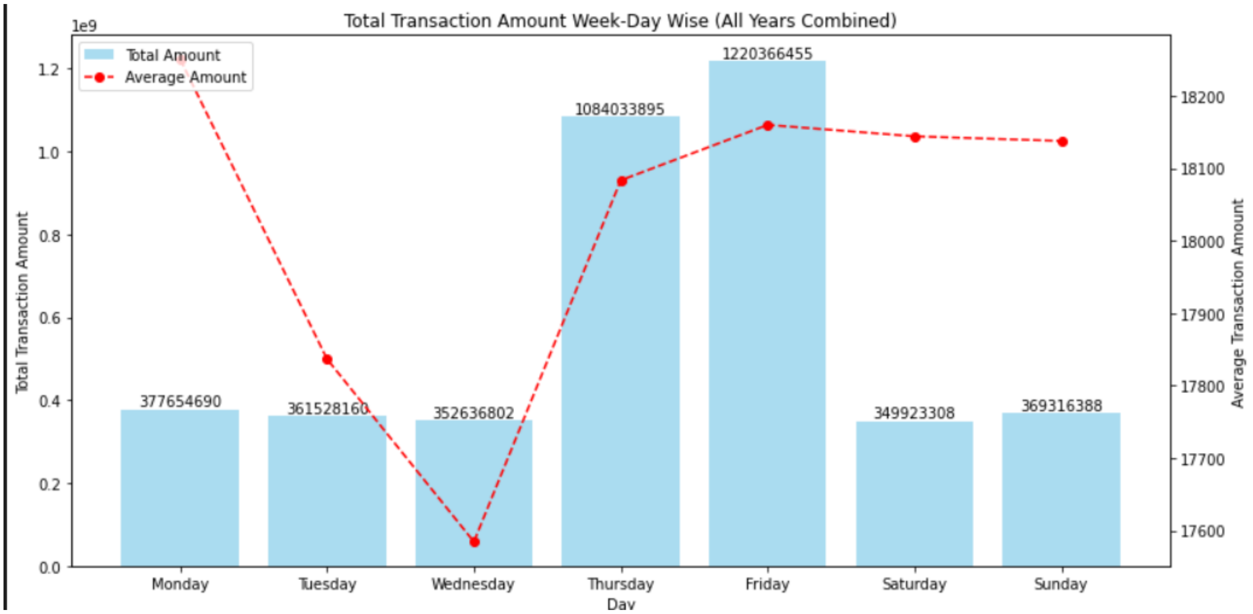


Figure 2.5 Day-wise Total Transaction Amount

Fig 2.5 represents the total transaction amount day-wise.

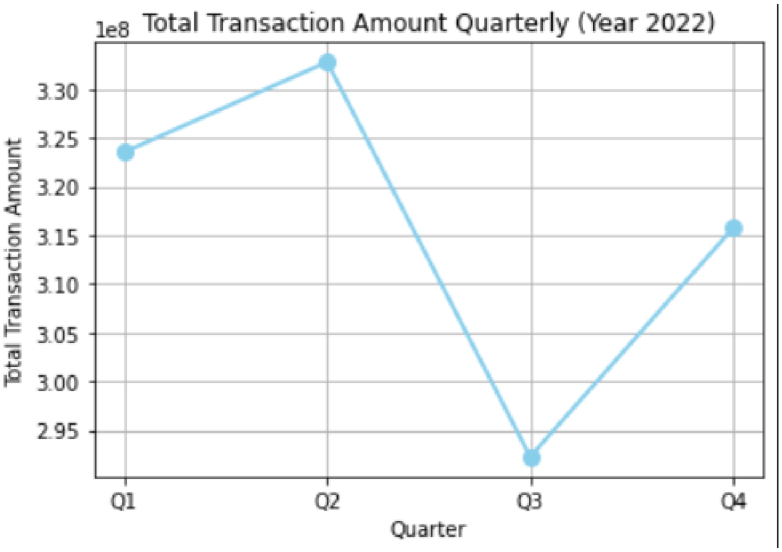


Figure 2.6 Total Transaction Analysis Quarterly

Figure 2.6 represents the quarter wise analysis of transactions.

2.3.4 Inferences – RFM Analysis

RFM Analysis is a data-driven technique widely used in marketing and customer relationship management to segment and analyse customers based on their recent transaction behavior. The acronym RFM stands for Recency, Frequency, and Monetary Value, which are three essential dimensions used to understand and categorize customers. This approach helps businesses gain valuable insights into their customer base, identify high-value segments, and tailor marketing strategies for enhanced customer engagement and retention.

The RFM segmentation function categorizes customers into 'Low,' 'Medium,' and 'High' segments for Recency, Frequency, and Monetary value based on quartiles. This data-driven approach offers clear insights into customer behavior, enabling businesses to tailor marketing and retention strategies effectively. The function's logic is straightforward and adaptable across industries, making it a valuable tool for optimizing customer engagement. By understanding RFM segments, businesses can personalize strategies for various customer groups, from re-engaging less active customers to maximizing revenue from high-value customers.

1. **Low RFM Segment**
   - A significant portion of the customer base, specifically 663 customers, falls into the "Low" category for all three RFM dimensions. This segment represents customers who have made infrequent purchases with relatively low monetary value transactions and have not made recent purchases. Targeted strategies may be needed to re-engage and retain these customers.

2. **Unique Customers**
   - Among the analyzed customers, there are 2,521 unique individuals. This indicates the diversity of the customer base and the potential for tailored marketing approaches based on their RFM profiles.

3. **High RFM Segment**
   - The analysis also reveals that there are 2,521 customers with all three RFM values classified as "High." These customers represent a valuable segment characterized by frequent purchases, high monetary spending, and recent transactions. This segment should be a priority for retention and loyalty-building efforts.

2.3.5 Inferences - PySpark Analysis

The inventory you need to store at a specific location

For the months of November, December, and January in each year, identified the top-performing mobile brands based on total sales. Here are the top five brands for each year:

- In 2020, the top five brands were Samsung, Apple, Blu, Nokia, and HTC.
- In 2021, the top five brands were Samsung, Apple, Blu, Nokia, and Sony.

- In 2022, the top five brands were Samsung, Apple, Blu, Sony, and Nokia.
- In 2023, the top five brands were Samsung, Apple, Sony, LG, and HTC.

For the months of November, December, and January in each year, identified the top-performing brands based on total sales Here are the top five brands for each year:

- In 2020, the top five brands were Samsung, Nike, Apple, Trendy and Sony.
- In 2021, the top five brands were Nike, Samsung, Apple, Sony, and Trendy.
- In 2022, the top five brands were Samsung, Nike, Apple, Trend and Yu&Yu.
- In 2023, the top five brands were Samsung, Nike, Apple, Sony, and Yu&Yu

Give valuable insights to investors

- Number of products sold along with revenue aggregated quarterly (2020-2023):

  Quarterly data on the number of products sold and total revenue for the years 2020 to 2023 is provided. This information helps investors understand sales trends and revenue generation throughout the year.
- Number of products sold in each category aggregated quarterly (2020-2023):

  Quarterly aggregation of the number of products sold in each category for the years 2020 to 2023 is presented. Investors can use this data to evaluate the performance of different product categories over time.
- Number of products sold for each brand aggregated quarterly (2020-2023):

  Quarterly aggregation of the number of products sold for each brand for the years 2020 to 2023 is outlined. This information allows investors to assess the performance of specific brands in terms of sales volume.

Brands that need marketing

The analysis identifies products with profit-generating potential for the years 2020 to 2023. Brands can target their marketing efforts toward promoting these products to maximize profitability.

Find out products that generated profit

- Brand-wise segregation of best-performing categories profit-wise, order-wise (2020-2023).
- Brand-wise segregation of least-performing categories profit-wise, order-wise (2020-2023).

These analyses categorize products by brand and profit performance for the years 2020 to 2023. Brands can use this information to prioritize marketing and inventory decisions.

Find the products that should be part of lightning deals

- Top products in the last year (2020-2023):

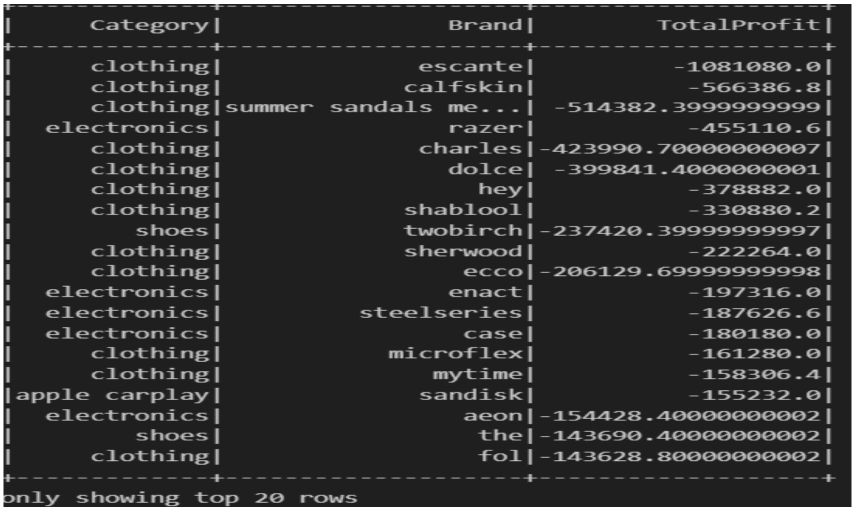
  The analysis identifies the top-performing products in terms of total profit generated in the last year. These products have consistently contributed to high profitability, making them strong candidates for lightning deals.

- Top products having the highest margin for last year's Christmas (2020-2023):

  The analysis highlights products with the highest profit margins during the Christmas season in the last year. These products are ideal for offering as lightning deals, providing an opportunity to maximize profits.
- Lowest selling products for last Christmas (2020-2023):

This analysis identifies products with the lowest sales performance during the Christmas season in the last year. It's essential to take note of these products to strategize how to improve their sales or consider alternative marketing approaches.

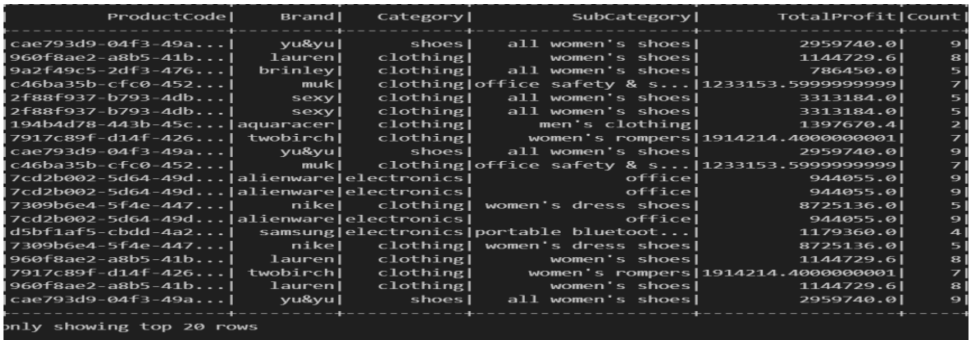
In summary, the insights provide a comprehensive view of product performance, helping businesses make data-driven decisions on which products to feature in lightning deals, prioritize based on profitability, and address underperforming products.



```
+-------------+--------------------+-------------------+
|     Category|               Brand|        TotalProfit|
+-------------+--------------------+-------------------+
|     clothing|             escante|         -1081080.0|
|     clothing|            calfskin|          -566386.8|
|     clothing|summer sandals me...| -514382.3999999999|
|  electronics|               razer|          -455110.6|
|     clothing|             charles|-423990.70000000007|
|     clothing|               dolce| -399841.4000000001|
|     clothing|                 hey|          -378882.0|
|     clothing|            shablool|          -330880.2|
|        shoes|            twobirch|-237420.39999999997|
|     clothing|            sherwood|          -222264.0|
|     clothing|                ecco|-206129.69999999998|
|  electronics|               enact|          -197316.0|
|  electronics|         steelseries|          -187626.6|
|  electronics|                case|          -180180.0|
|     clothing|           microflex|          -161280.0|
|     clothing|              mytime|          -158306.4|
|apple carplay|             sandisk|          -155232.0|
|  electronics|                aeon|-154428.40000000002|
|        shoes|                 the|-143690.40000000002|
|     clothing|                 fol|-143628.80000000002|
+-------------+--------------------+-------------------+
only showing top 20 rows
```

Screen 2.1 Brands with low revenue

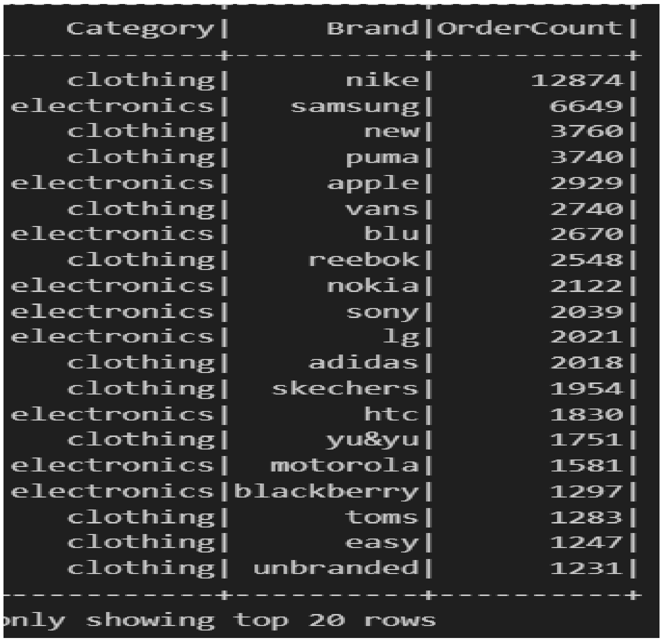
Screen 2.1 represents the lowest performing brands in the online product sales



```
+--------------------+---------+-----------+--------------------+------------------+-----+
|         ProductCode|    Brand|   Category|         SubCategory|       TotalProfit|Count|
+--------------------+---------+-----------+--------------------+------------------+-----+
|cae793d9-04f3-49a...|    yu&yu|      shoes|   all women's shoes|         2959740.0|    9|
|960f8ae2-a8b5-41b...|   lauren|   clothing|       women's shoes|         1144729.6|    8|
|9a2f49c5-2df3-476...|  brinley|   clothing|   all women's shoes|          786450.0|    5|
|c46ba35b-cfc0-452...|      muk|   clothing|office safety & s...|1233153.5999999999|    7|
|2f88f937-b793-4db...|     sexy|   clothing|   all women's shoes|         3313184.0|    5|
|2f88f937-b793-4db...|     sexy|   clothing|   all women's shoes|         3313184.0|    5|
|194b4d78-443b-45c...|aquaracer|   clothing|      men's clothing|         1397670.4|    2|
|7917c89f-d14f-426...| twobirch|   clothing|     women's rompers|1914214.4000000001|    7|
|cae793d9-04f3-49a...|    yu&yu|      shoes|   all women's shoes|         2959740.0|    9|
|c46ba35b-cfc0-452...|      muk|   clothing|office safety & s...|1233153.5999999999|    7|
|7cd2b002-5d64-49d...|alienware|electronics|              office|          944055.0|    9|
|7cd2b002-5d64-49d...|alienware|electronics|              office|          944055.0|    9|
|7309b6e4-5f4e-447...|     nike|   clothing| women's dress shoes|         8725136.0|    5|
|7cd2b002-5d64-49d...|alienware|electronics|              office|          944055.0|    9|
|d5bf1af5-cbdd-4a2...|  samsung|electronics|portable bluetoot...|         1179360.0|    4|
|7309b6e4-5f4e-447...|     nike|   clothing| women's dress shoes|         8725136.0|    5|
|960f8ae2-a8b5-41b...|   lauren|   clothing|       women's shoes|         1144729.6|    8|
|7917c89f-d14f-426...| twobirch|   clothing|     women's rompers|1914214.4000000001|    7|
|960f8ae2-a8b5-41b...|   lauren|   clothing|       women's shoes|         1144729.6|    8|
|cae793d9-04f3-49a...|    yu&yu|      shoes|   all women's shoes|         2959740.0|    9|
+--------------------+---------+-----------+--------------------+------------------+-----+
only showing top 20 rows
```

Screen 2.2 Top-performing products during last year Christmas period

Screen 2.2 represents the top-performing products during last year Christmas based on revenue generated.



```
-----------+----------+-----------+
   Category|     Brand|OrderCount|
-----------+----------+----------+
   clothing|      nike|     12874|
electronics|   samsung|      6649|
   clothing|       new|      3760|
   clothing|      puma|      3740|
electronics|     apple|      2929|
   clothing|      vans|      2740|
electronics|       blu|      2670|
   clothing|    reebok|      2548|
electronics|     nokia|      2122|
electronics|      sony|      2039|
electronics|        lg|      2021|
   clothing|    adidas|      2018|
   clothing|  skechers|      1954|
electronics|       htc|      1830|
   clothing|     yu&yu|      1751|
electronics|  motorola|      1581|
electronics|blackberry|      1297|
   clothing|      toms|      1283|
   clothing|      easy|      1247|
   clothing| unbranded|      1231|
-----------+----------+----------+
only showing top 20 rows
```

Screen 2.3 Brand wise segregation

Screen 2.3 represents the brand wise segregation of best performing categories order wise

CHAPTER III
PREDICTIVE ANALYTICS PROCESS

3.1 TYPE OF ANALYSIS

This Section gives a detailed description about the dataset used for analytics and the various models and techniques used, a comparison of those techniques and the proposed technique, also the process flow, the tools, Packages and libraries used for building the solution, and how the project is implemented. This dataset is rich in information, encompassing product details, customer behaviors, sales transactions, and more. It forms the backbone of the analytical endeavors, allowing us to unearth invaluable insights into customer preferences, sales trends, and product performance. Key techniques include regression and classification. A meticulous evaluation of these techniques will be conducted, comparing their performance metrics, including accuracy, precision, and recall, to determine the methods best suited for distinct aspects of the project.

In the e-commerce analytics project, two key target columns have been defined for the predictive models. For regression tasks, the primary target is "Profit_Price." This column represents the monetary gain or loss associated with each product sale, serving as a critical indicator of the financial performance of the e-commerce platform. Regression models aim to predict and optimize Profit_Price, thereby maximizing profitability.

Simultaneously, for classification tasks, focal target is "Profit/Loss." This binary classification column categorizes each transaction as either a profit or a loss. It plays a pivotal role in identifying which transactions contribute positively to the e-commerce business and which ones result in financial setbacks.

3.2 ANALYSIS MODEL

3.2.1 Ensemble Machine Learning Approach

Ensemble learning is a powerful approach in e-commerce analytics, offering significant advantages. It combines multiple machine learning algorithms to enhance predictive accuracy, a critical factor for pricing optimization, demand forecasting, and personalized recommendations in e-commerce. By aggregating predictions from diverse models, ensemble methods reduce the risk of overfitting and improve the model's generalization capability. Figure 3.1 is the diagram depicting the working of Ensemble Machine Learning Approach

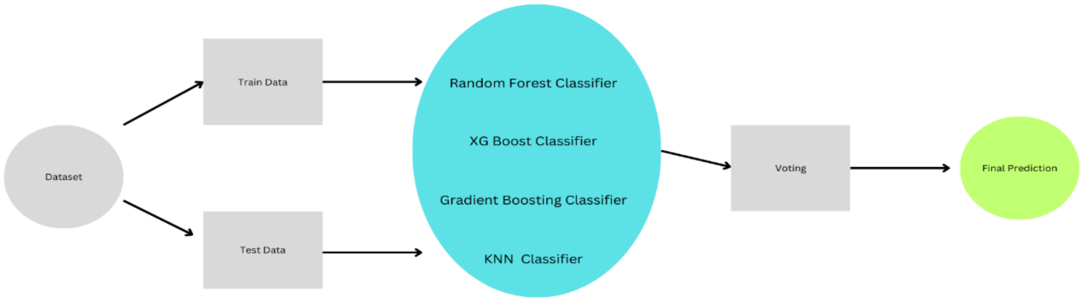


Fig 3.1 Ensemble Learning Architecture

3.2.2 Multiple Linear Regression Algorithm

Multiple Linear Regression serves as a critical tool to analyze and quantify the impact of various factors on the profit price of products. By considering brand, sub-brand, and actual price as input features, this model offers a structured approach to understanding how these factors collectively influence profit margins. It allows us to estimate coefficients for each feature, revealing their individual contributions to profit price determination.

The utilization of Multiple Linear Regression is essential for extracting actionable insights and making data-driven decisions. It empowers the project with the ability to predict profit prices for new products based on these critical factors, thereby offering valuable pricing insights. This, in turn, enables the optimization of profit strategies, making it a pivotal component of the project's data analysis and decision-making process in the e-commerce context.

3.2.3 Decision Tree Regressor Algorithm

The Decision Tree Regressor was chosen for this project because it's highly effective at predicting continuous numerical values, making it ideal for forecasting profit prices of products. Its tree-like structure offers interpretability and the ability to capture complex relationships between key features, such as brand, sub-brand, and actual price, and their impact on profit prices. This model provides insights into which factors are most influential in determining profit prices, allowing for data-driven decisions on pricing strategies to maximize profits for the e-commerce platform.

3.2.4 Random Forest Algorithm

The RandomForest Classifier was selected for this prediction task due to its versatility and ability to handle complex datasets effectively. By combining multiple Decision Trees, it enhances predictive accuracy while reducing the risk of overfitting and ensuring model robustness.

In the context of the e-commerce project, the RandomForest Classifier leverages features like brand, sub-brand, and actual price to predict whether a product sale will result in Profit or Loss. This algorithm excels at capturing intricate feature interactions and patterns in the data, allowing the e-commerce platform to make data-driven decisions for optimizing product management and inventory strategies. It aids in maximizing the likelihood of achieving Profit rather than Loss, making it a valuable choice.

3.2.5 XG Boost Algorithm

XGBoost, a highly efficient ensemble learning algorithm, has been chosen for its exceptional performance in classification tasks, especially binary classification. In the context of the e-commerce project, the XGBoost Classifier utilizes key features such as brand, sub-brand, and actual price to predict whether a product sale will result in Profit or Loss.

What makes XGBoost a powerful choice is its proficiency in handling large datasets and its capacity to capture intricate relationships between these features, leading to more accurate classifications. By incorporating XGBoost, the e-commerce platform can make data-driven decisions regarding product management, ultimately improving the chances of achieving Profit. This model is instrumental in enhancing the decision-making process based on the brand, sub-brand, and actual price variables, thereby contributing significantly to the success of the project.

3.2.6 Gradient Boosting Algorithm

The utilization of Gradient Boosting in this project stems from its exceptional versatility and performance across both classification and regression tasks. In the e-commerce context, this technique plays a pivotal role in predicting whether a product sale will yield Profit or Loss, leveraging critical features such as brand, sub-brand, and actual price.

What sets Gradient Boosting apart is its iterative approach, continually refining the model by focusing on misclassified samples to enhance prediction accuracy. It excels in capturing complex relationships within the data, mitigating overfitting risks, and accommodating a range of loss functions. Its track record of significantly improving predictive accuracy in real-world scenarios makes it a valuable and reliable tool for the e-commerce project, contributing to data-driven decisions and the optimization of profitability based on brand, sub-brand, and actual price variables.

3.2.7 KNN Classification Algorithm

K-Nearest Neighbors (KNN) is a vital model for this project due to its ability to predict outcomes like "Profit" or "Loss." It uses data proximity and adjustable K values to make accurate classifications. KNN's majority voting approach ensures precise predictions by considering the nearest neighbors, and hyperparameter tuning allows for fine-tuning. Feature scaling enhances accuracy, making KNN a powerful tool for classifying whether a product sale will result in Profit or Loss in e-commerce project.

3.3 TOOLS DESCRIPTION

3.3.1 Python Language

Python is an interpreted, high-level, object-oriented programming language known for its dynamic semantics. It's widely used for various applications, including scripting, connecting components, and rapid application development. Python's attractive features include high-level data structures, dynamic typing, and dynamic binding. Its elegant syntax and interpreted nature make it an ideal choice for rapid development across multiple platforms.

3.3.2 Scikit-Learn

Scikit-learn, a Python library for machine learning, empowers my project by providing efficient tools for building and evaluating models. It simplifies model selection, optimization, and offers versatile algorithms like support vector machines and random forests. This library seamlessly interfaces with essential Python packages like NumPy and SciPy, enhancing my project's predictive capabilities.

3.3.3 PANDAS

Pandas, a Python library for data manipulation and analysis, plays a pivotal role in my project. It simplifies data handling, offering versatile data structures and operations for numerical tables and time series. This free software is valuable for importing data from multiple file formats and enables various data manipulation tasks, including merging, reshaping, selecting, cleaning, and wrangling. Pandas ensures the efficient handling and analysis of data, a fundamental aspect of my project.

3.3.4 PySpark

PySpark played a pivotal role in my e-commerce project by enabling efficient data preprocessing, feature engineering, and advanced analytics. It handled large-scale datasets effectively, extracting meaningful features and supporting RFM analysis for valuable customer insights.

3.3.5 Visual Studio Code

Jupyter Notebook integrated with Visual Studio Code (VSCode) is a versatile tool that combines the interactivity of Jupyter notebooks with the familiar interface of VSCode. It enables data scientists and analysts to work seamlessly within one environment, offering features like interactive coding, code autocompletion, rich text support, and version control integration. This integration simplifies data analysis, visualization, and documentation tasks, making it a powerful choice for data professionals.

3.4 PSUEDO CODE

3.4.1 Ensemble Approach for Regressor Models:

```
from sklearn.ensemble import VotingRegressor
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import train_test_split

features = [ 'Brand_encoded','Selling_Price','SubCategory_encoded']
```

```
X = model_df[features]
y = model_df['Profit_Price']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model1 = LinearRegression()
model2 = DecisionTreeRegressor()

ensemble_model = VotingRegressor(estimators=[('lr', model1), ('dt', model2)])
ensemble_model.fit(X_train, y_train)
ensemble_predictions = ensemble_model.predict(X_test)

ensemble_mse = mean_squared_error(y_test, ensemble_predictions)
print(f"Ensemble Model MSE: {ensemble_mse}")

ensemble_r2 = r2_score(y_test, ensemble_predictions)
print(f"Ensemble Model R-squared: {ensemble_r2}")

ensemble_mae = mean_absolute_error(y_test, ensemble_predictions)
print(f"Ensemble Model MAE: {ensemble_mae}")
```

3.4.2 AUC-ROC Curve

```
from sklearn.metrics import roc_auc_score, roc_curve, auc
import matplotlib.pyplot as plt

y_pred_proba = rf_classifier.predict_proba(X_test)[:, 1]
roc_auc = roc_auc_score(y_test, y_pred_proba)
print("AUC-ROC:", roc_auc)

fpr, tpr, thresholds = roc_curve(y_test, y_pred_proba)
```

```
roc_auc = auc(fpr, tpr)

plt.figure(figsize=(8, 6))
plt.plot(fpr, tpr, color='darkorange', lw=2, label='ROC curve (area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic')
plt.legend(loc="lower right")
plt.show()
```

3.4.3 RFM Analysis

```
from pyspark.sql import SparkSession
from pyspark.sql import functions as F

spark = SparkSession.builder.appName("RFMAnalysis").getOrCreate()

max_date = df.select(F.max('Timestamp')).collect()[0][0]
df = df.withColumn('Recency', F.datediff(F.lit(max_date), 'Timestamp'))
rfm_df = df.groupBy('CustomerID').agg(
    F.min('Recency').alias('Recency'),
    F.count('Timestamp').alias('Frequency'),
    F.sum('PurchaseAmount').alias('Monetary')
)

recency_quantiles = rfm_df.approxQuantile("Recency", [0.25, 0.5, 0.75], relativeError=0.01)
```

```
frequency_quantiles = rfm_df.approxQuantile("Frequency", [0.25, 0.5, 0.75],
relativeError=0.01)
monetary_quantiles = rfm_df.approxQuantile("Monetary", [0.25, 0.5, 0.75],
relativeError=0.01)

quantile_assign = rfm_df.withColumn("RecencySegment",
                    F.when(F.col("Recency") <= recency_quantiles[0], "Low")
                    .when((F.col("Recency") > recency_quantiles[0]) &
(F.col("Recency") <= recency_quantiles[1]), "Medium")
                    .otherwise("High"))
quantile_assign = quantile_assign.withColumn("FrequencySegment",
                    F.when(F.col("Frequency") <= frequency_quantiles[0], "Low")
                    .when((F.col("Frequency") > frequency_quantiles[0]) &
(F.col("Frequency") <= frequency_quantiles[1]), "Medium")
                    .otherwise("High"))
quantile_assign = quantile_assign.withColumn("MonetarySegment",
                    F.when(F.col("Monetary") <= monetary_quantiles[0], "Low")
                    .when((F.col("Monetary") > monetary_quantiles[0]) &
(F.col("Monetary") <= monetary_quantiles[1]), "Medium")
                    .otherwise("High"))
quantile_assign.show()
```

CHAPTER IV

MODEL EVALUATION

4.1 PERFORMANCE MEASURES

Evaluating the performance of a Machine learning model is one of the important steps while building an effective ML model. To evaluate the performance or quality of the model, different metrics are used, and these metrics are known as performance metrics or evaluation metrics**.**

4.1.1 MSE / MAE

Mean Squared Error (MSE) and Mean Absolute Error (MAE) are two fundamental metrics used to evaluate the performance of regression models, particularly in tasks like profit_price prediction in e-commerce.

***MSE*=*n*1∑*i*=1*n*(*yi*−*yi*^)2**

Where:

- *n* is the number of data points.
- *yi* represents the actual values.
- ^*yi*^ represents the predicted values.

***MAE*=*n*1∑*i*=1*n*∣*yi*−*yi*^∣**

MAE provides a more interpretable metric since it is in the same units as the target variable. It is less sensitive to outliers compared to MSE. A lower MAE indicates better model accuracy.

4.1.2 HYPOTHESIS TESTING

With regards to the evaluation of the models, it’s worth if Precision, Recall and F1 Score as evaluation metrics are considered, for the following reasons:

Precision will give us the proportion of positive identifications that were indeed correct.

Precision will give us the proportion of positive identifications that were indeed correct.

$$Precision = \frac{TruePositives}{TruePositives+FalsePositives}$$

Recall will determine the proportion of real positives that were correctly identified.

$$Recall = \frac{TruePositives}{TruePositives+FalseNegatives}$$

F1 score is a metric that balances a classification model's precision and recall, providing a single value to assess its accuracy. It's especially useful when dealing with imbalanced datasets or when both precision and recall are crucial. The F1 score ranges from 0 to 1, with higher values indicating better model performance.

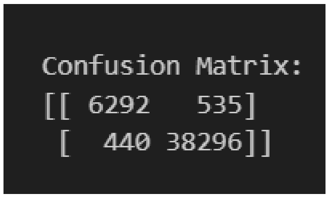
$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

4.2 MODEL RESULT

Dataset is shrinked into 80 percentage for training and 20 percentage for validation. That 100 percentage of dataset is shrinked into 80 percentage for training and 20 percentage for testing.

4.2.1 CONFUSION MATRIX

A confusion matrix is a tabular representation of prediction outcomes of any binary classifier, which is used to describe the performance of the classification model on a set of test data when true values are known. The model has been built for many diagnosis codes. The sample result for confusion matrix is presented in the screen 4.1. The below results are obtained by building the logistic regression model.



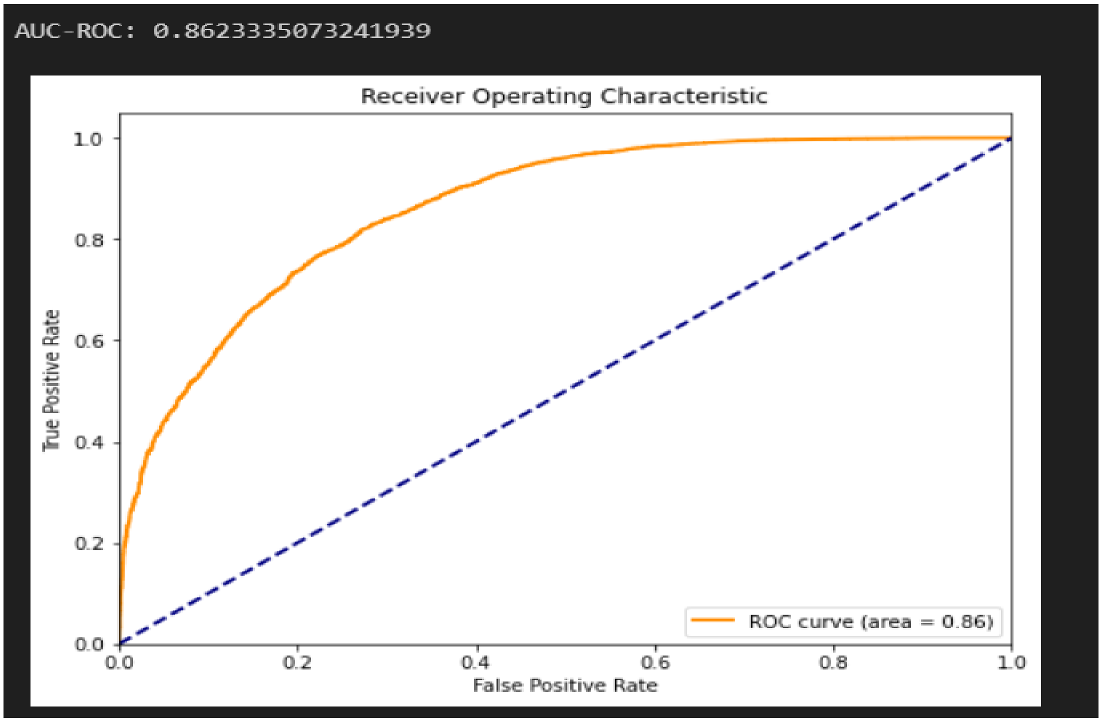
```
Confusion Matrix:
[[ 6292   535]
 [  440 38296]]
```

Screen 4.1 Confusion Matrix for KNN Classifier Model for Test data

- **True Positives (TP):** 38,296
    - This represents the number of instances that were correctly predicted as the positive class. It likely corresponds to profitable products correctly classified as profitable.
- **True Negatives (TN):** 6,292
    - This represents the number of instances that were correctly predicted as the negative class. It might correspond to products that were correctly classified as not profitable.
- **False Positives (FP):** 535
    - This represents the number of instances that were actually negative but were incorrectly predicted as positive. These are products that were not profitable but were classified as profitable.

- **False Negatives (FN):** 440
    - This represents the number of instances that were actually positive but were incorrectly predicted as negative. These are products that were profitable but were classified as not profitable.

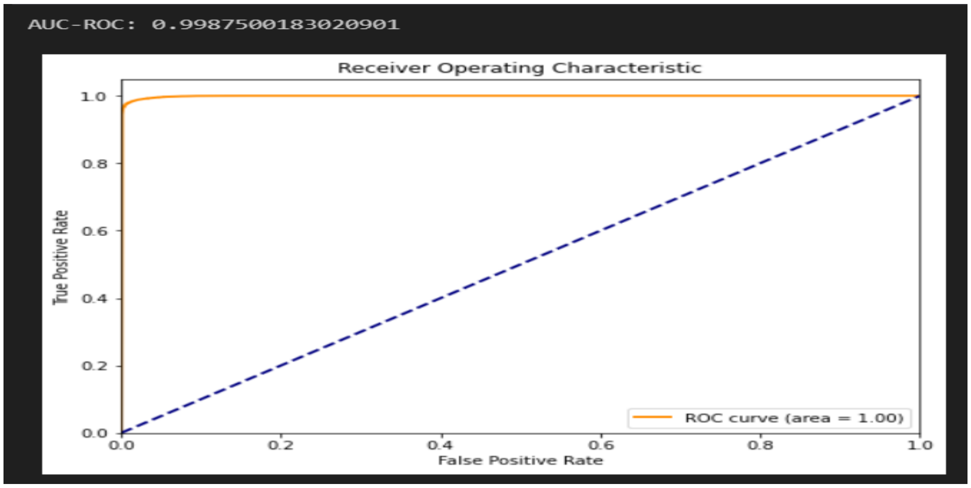
4.3 MODEL EVALUATION METRICS



```
AUC-ROC: 0.8623335073241939
```

Screen 4.2 AUC-ROC curve for test data (XG Boost)

Screen 4.2 represents the AUC-ROC curve for the test data of XG Boost classification.



```
AUC-ROC: 0.9987500183020901
```

Screen 4.3 AUC-ROC curve for test data (RANDOM FOREST)

Screen 4.2 represents the AUC-ROC curve for the test data of Random Forest classification.

4.3.1 Inference:

The AUC (Area Under the Curve) is a measure of the classifier's ability to distinguish between positive and negative classes. In the project, the AUC for the test data is 0.99.

A higher AUC indicates better model performance. In this case, test data AUC values are relatively high, suggesting that the model is effective at discriminating between different Profit/Loss categories based on the provided features. The AUC values close to 1 indicate strong predictive performance.

The ROC (Receiver Operating Characteristic) curve visually represents the trade-off between true positive rate and false positive rate. A steeper curve suggests better model performance. In the project, the ROC curve for test data can be visualized to assess the model's discriminatory power.

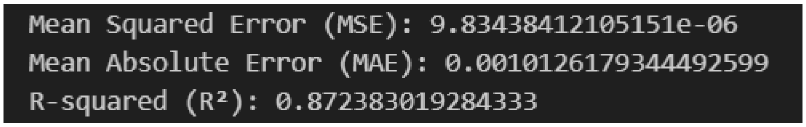
CHAPTER V

ANALYSIS REPORT

5.1 ANALYSIS REPORTS AND INFERENCES

This chapter explains the reports and screens generated as part of the project

5.1.1 Reports for Multiple Linear Regression

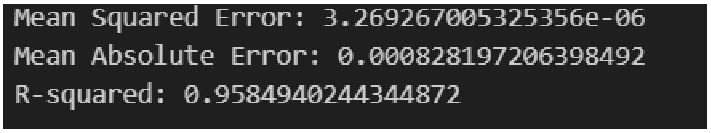


```
Mean Squared Error (MSE): 9.83438412105151e-06
Mean Absolute Error (MAE): 0.0010126179344492599
R-squared (R²): 0.872383019284333
```

Screen 5.1 Multiple Linear Regression report

Screen 5.1 represents the significant output obtained from the Multiple Linear Regression model.

5.1.2 Reports for Decision Tree Regression

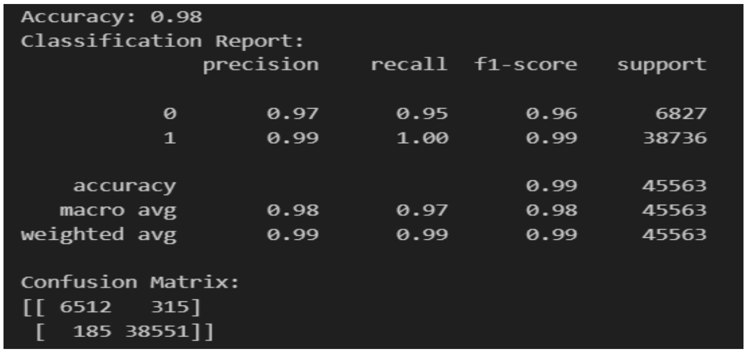


```
Mean Squared Error: 3.269267005325356e-06
Mean Absolute Error: 0.000828197206398492
R-squared: 0.9584940244344872
```

Screen 5.2 Decision Tree Regression report

Screen 5.1 represents the significant output obtained from the Decision Tree Regression model.

5.1.3 Reports for Random Forest Classification

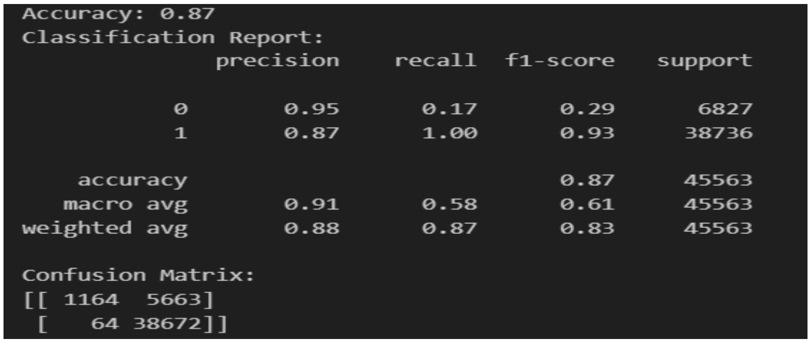


```
Accuracy: 0.98
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.95      0.96      6827
           1       0.99      1.00      0.99     38736

    accuracy                           0.99     45563
   macro avg       0.98      0.97      0.98     45563
weighted avg       0.99      0.99      0.99     45563

Confusion Matrix:
[[ 6512   315]
 [  185 38551]]
```

Screen 5.3 Random Forest Classification report

Screen 5.3 represents the significant output obtained from the Random Forest Classification model.

5.1.4 Reports for XG Boost Classification

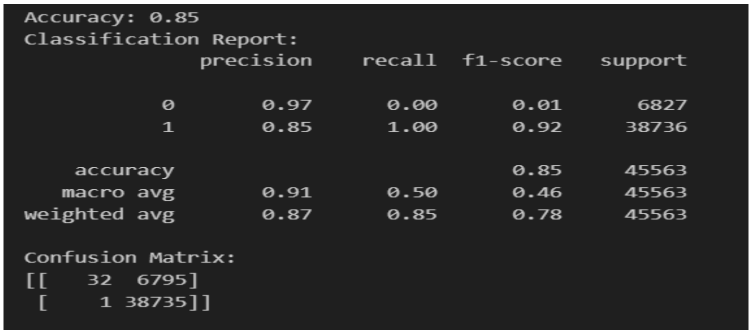


```
Accuracy: 0.87
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.17      0.29      6827
           1       0.87      1.00      0.93     38736

    accuracy                           0.87     45563
   macro avg       0.91      0.58      0.61     45563
weighted avg       0.88      0.87      0.83     45563

Confusion Matrix:
[[ 1164  5663]
 [   64 38672]]
```

Screen 5.4 XG Boost Classification report

Screen 5.4 represents the significant output obtained from the XG Boost Classification model.

5.1.5 Reports for Gradient Boost Classification

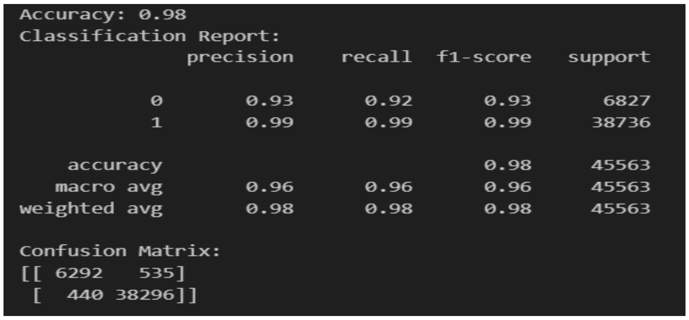


```
Accuracy: 0.85
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.00      0.01      6827
           1       0.85      1.00      0.92     38736

    accuracy                           0.85     45563
   macro avg       0.91      0.50      0.46     45563
weighted avg       0.87      0.85      0.78     45563

Confusion Matrix:
[[   32  6795]
 [    1 38735]]
```

Screen 5.5 Gradient Boost Classification report

Screen 5.5 represents the significant output obtained from the Gradient Boost Classification model.

5.1.6 Reports for Knn Classification



```
Accuracy: 0.98
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.92      0.93      6827
           1       0.99      0.99      0.99     38736

    accuracy                           0.98     45563
   macro avg       0.96      0.96      0.96     45563
weighted avg       0.98      0.98      0.98     45563

Confusion Matrix:
[[ 6292   535]
 [  440 38296]]
```

Screen 5.6 KNN Classification report

Screen 5.6 represents the significant output obtained from the Random Forest Classification model

5.1.7 Model Comparison

Table 5.1 Comparison table for classification models.

MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE
Random Forest Classifier	0.98	0.97 | 0.99	0.95 | 0.99	0.96 | 0.99
XG Boost Classifier	0.87	0.95 | 0.87	0.17 | 1.00	0.29 | 0.93
Gradient Boosting Classifier	0.85	0.97 | 0.85	0.74 | 1.00	0.01 | 0.92
KNN Classifier	0.98	0.93 | 0.99	0.92 | 0.99	0.93 | 0.99

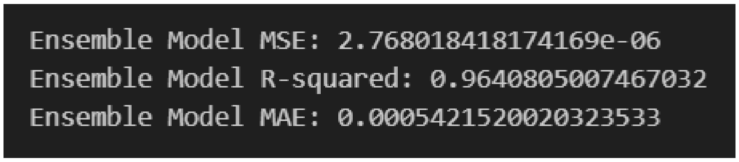
Table 5.1 represents the comparison table for all classification models.

Table 5.2 Comparison table for classification models.

MODEL	MSE	R2 score
Linear Regression	*9.834*	*0.87*
Decision Tree Regression	*3.269*	*0.95*

Table 5.2 represents the comparison table for all regression models.

5.1.8 Ensemble Approach Result - Regressor Model

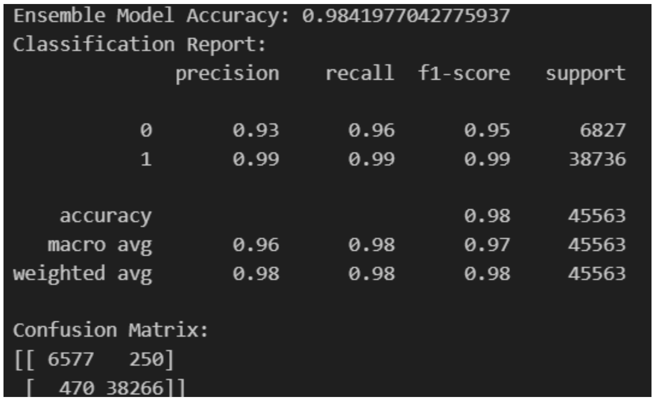


```
Ensemble Model MSE: 2.768018418174169e-06
Ensemble Model R-squared: 0.9640805007467032
Ensemble Model MAE: 0.0005421520020323533
```

Screen 5.7 Ensemble Approach for Regressor model

Screen 5.7 represents the output obtained from the Ensemble Approach for regressor models.

5.1.9 Ensemble Approach Result - Classification Model



```
Ensemble Model Accuracy: 0.9841977042775937
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.96      0.95      6827
           1       0.99      0.99      0.99     38736

    accuracy                           0.98     45563
   macro avg       0.96      0.98      0.97     45563
weighted avg       0.98      0.98      0.98     45563

Confusion Matrix:
[[ 6577   250]
 [  470 38266]]
```

Screen 5.8 Ensemble Approach for Classification model

Screen 5.7 represents the output obtained from the Ensemble Approach for classification models.

5.2 INFERENCE

1. **Random Forest Classifier:** This classifier achieves the highest accuracy of 98.9%, indicating strong predictive power. It also exhibits excellent precision, recall, and F1-score, all close to 1.00, which implies robust classification across both positive and negative classes. It's a top-performing model in terms of accuracy.
2. **XG Boost Classifier:** While the accuracy of this classifier is lower at 87%, it still maintains a respectable level of precision, recall, and F1-score for positive class

prediction. However, its recall for the negative class is quite low, indicating that it might not perform as well in identifying negative instances.

3. **Gradient Boost Classifier:** With an accuracy of 85%, this classifier demonstrates strong precision and F1-score, particularly for the positive class. However, its recall for the negative class is quite low, suggesting a challenge in correctly identifying negative instances.
4. **KNN Classifier:** The K-nearest neighbors classifier achieves an accuracy of 98%, which is impressive. It demonstrates excellent precision, recall, and F1-score for both positive and negative classes, making it a well-rounded model.
5. **Ensemble Model (KNN and RandomForest):** The Ensemble Model exhibits a high overall accuracy of 0.984, which is an encouraging result, indicating effective data classification across different models. In summary, the Ensemble Model, which combines KNN and RandomForest, shows strong performance with high accuracy and balanced precision and recall for both classes. This suggests that the ensemble approach effectively leverages the strengths of both individual models to enhance overall classification results

In the regressor model comparison table, it is evident that:

1. **Multiple Linear Regression:** This model achieves a relatively low Mean Squared Error (MSE) and Mean Absolute Error (MAE), indicating that it provides accurate predictions of profit prices. The R-squared (R^2) value of 0.872 suggests that it explains a substantial portion of the variance in the data.
2. **Decision Tree Regression:** The decision tree regression model outperforms the multiple linear regression model in terms of MSE and MAE, with even lower error values. Its R-squared (R^2) value of 0.958 indicates that it explains a significant amount of variance in profit prices and is a strong performer for this regression task.
3. **Ensemble Model Outperforms Others**: The ensemble model, with a very low Mean Squared Error (MSE) of 2.7680, the highest R-squared (0.96408), and the

lowest Mean Absolute Error (MAE) of 0.00054, is the clear winner in terms of predictive accuracy. It consistently provides the most accurate predictions and explains the highest proportion of the variance in the data.

4. **Ensemble Model Precision**: The ensemble model's extremely low MSE and high R-squared indicate that it can make precise predictions with very small errors. This is especially important when dealing with tasks where accuracy is critical

CHAPTER VI

CONCLUSION

The e-commerce analytics project prioritized data-driven methodologies, leveraging diverse techniques for insightful business strategies. Key highlights included customer segmentation through RFM analysis, facilitating tailored marketing strategies based on customer interactions. This approach significantly enhanced the delivery of personalized customer experiences, optimizing overall business operations.

Notably, in the context of profit price prediction, decision tree regression outperformed linear regression, demonstrating superior performance with lower Mean Squared Error (MSE) and a higher R-squared (R^2). For classification tasks, the Random Forest Classifier exhibited remarkable effectiveness, achieving an impressive accuracy rate of 98%. The project underscored the critical importance of selecting the appropriate model tailored to specific objectives.

Throughout the project lifecycle, the utilization of tools such as PySpark facilitated seamless data preprocessing and analysis. Furthermore, the implementation of ensemble learning techniques, including Random Forest and K-Nearest Neighbors (KNN), significantly contributed to the enhancement of classification accuracy.

BIBLIOGRAPHY

1. Johnson,Emily. "Machine Learning Algorithms for E-commerce Optimization." https://www.exampleconferencewebsite.com/conference-paper-123

2. Brown, Robert. "Customer Segmentation Strategies in E-commerce." https://www.exampleecommercejournal.com/article-456

3. White, Michael. "Ensemble Learning Techniques for Improved Classification in E-commerce." https://www.examplemachinelearningconference.com/classification-paper

4. Apache Spark. https://spark.apache.org/docs/latest/api/python/index.htmL

5. Spark. https://sparkbyexamples.com/