K-mart sales 2019   
Data Analysis

Aziz Kerimzhanov  
AJ23SYD004

Contents

[Abstract: 3](#_Toc135605900)

[Introduction: 3](#_Toc135605901)

[Data description: 4](#_Toc135605902)

[Exploratory Data Analysis (EDA) 6](#_Toc135605903)

[During which month did sales peak? 6](#_Toc135605904)

[In which city was the largest number of sales recorded? 6](#_Toc135605905)

[What product sold the most? 7](#_Toc135605906)

[The most appropriate time to display ads to increase the likelihood of buyers buying goods? 7](#_Toc135605907)

[What products are most often sold together? 9](#_Toc135605908)

[Product wise price 10](#_Toc135605909)

[Machine learn technique – Classification model 10](#_Toc135605910)

[Classification model 11](#_Toc135605911)

[Conclusion: 12](#_Toc135605912)

[Recommendation 14](#_Toc135605913)

[References: 15](#_Toc135605914)

# Abstract:

This data analysis report was made for exploring the K-mart sales in 2019 database to find valuable insights and answer to several business questions. The report includes a comprehensive exploratory data analysis of sales data, products, and customer behavior. Using different techniques, the report provides with meaningful information that can help to K-mart team to develop the sales strategy for the next years.

The dataset contains several key columns, including “Order ID”, “Product”, “Quantity Ordered”, “Price Each”, “Order Date”, “Purchase Address”, “month”. These columns contain important information about sales transactions, product information and customer orders.

The report begins with database introducing and describing the purpose of the report. It provides the data structure, EDA analysis and Classification model technique.

Introduction:  
I am a student of S P Jain Global Management school that is located in Australia, Sydney. In Sydney, one of the popular retail stores is K-mart. I often visit this store because it offers a wide range of products at affordable prices. I like this retail and I was interested if there is a dataset of that market. After conducting research on the internet, I came across sales data for the year 2019, but it pertained to the US market. Nevertheless, I decided to analyze that dataset to uncover valuable insights and patterns.

# Data description:

The K-mart sales for 2019 contains 12 datasets for each month:  
“Sales\_January\_2019”  
“Sales\_February\_2019”

“Sales\_March\_2019”  
“Sales\_April\_2019”  
“Sales\_May\_2019”  
“Sales\_June\_2019”  
“Sales\_July\_2019”  
“Sales\_August\_2019”

“Sales\_September\_2019”

“Sales\_October\_2019”

“Sales\_November\_2019”  
“Sales\_December\_2019”

Each dataset contains 6 columns:  
“Order ID”, “Product”, “Quantity Ordered”, “Price Each”, “Order Date”, “Purchase Address”

And from 11000 till 25000 rows

The shape of dataset is:  
(9723, 7)

(12036, 7)

(15226, 7)

(18383, 7)

(16635, 7)

(13622, 7)

(14371, 7)

(12011, 7)

(11686, 7)

(20379, 7)

(17661, 7)

(25117, 7)

All columns contains object data type  
Order ID object

Product object

Quantity Ordered object

Price Each object

Order Date object

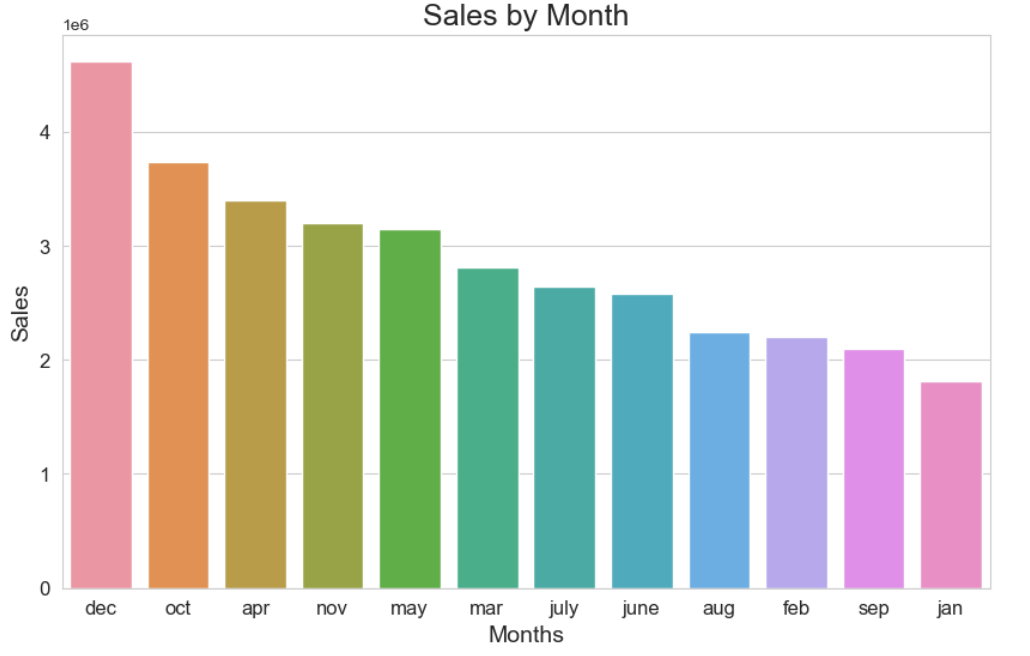
Purchase Address object

month object

Unique Product — 19  
Total Orders — 178437  
Cities — 9

# Exploratory Data Analysis (EDA)

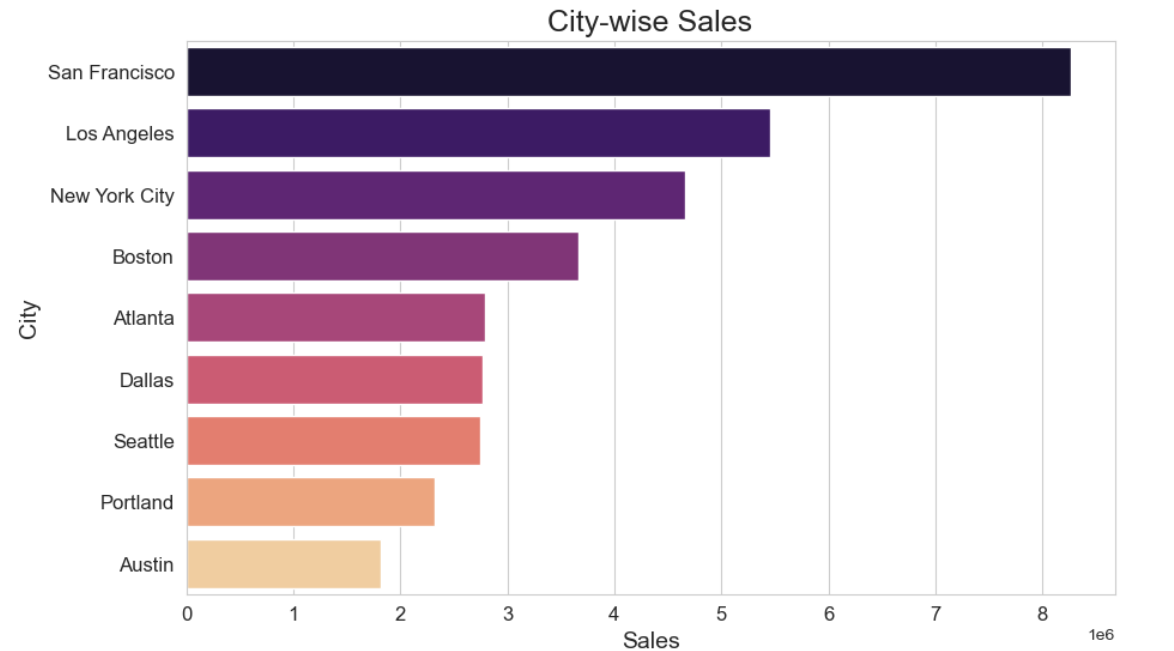
## During which month did sales peak?



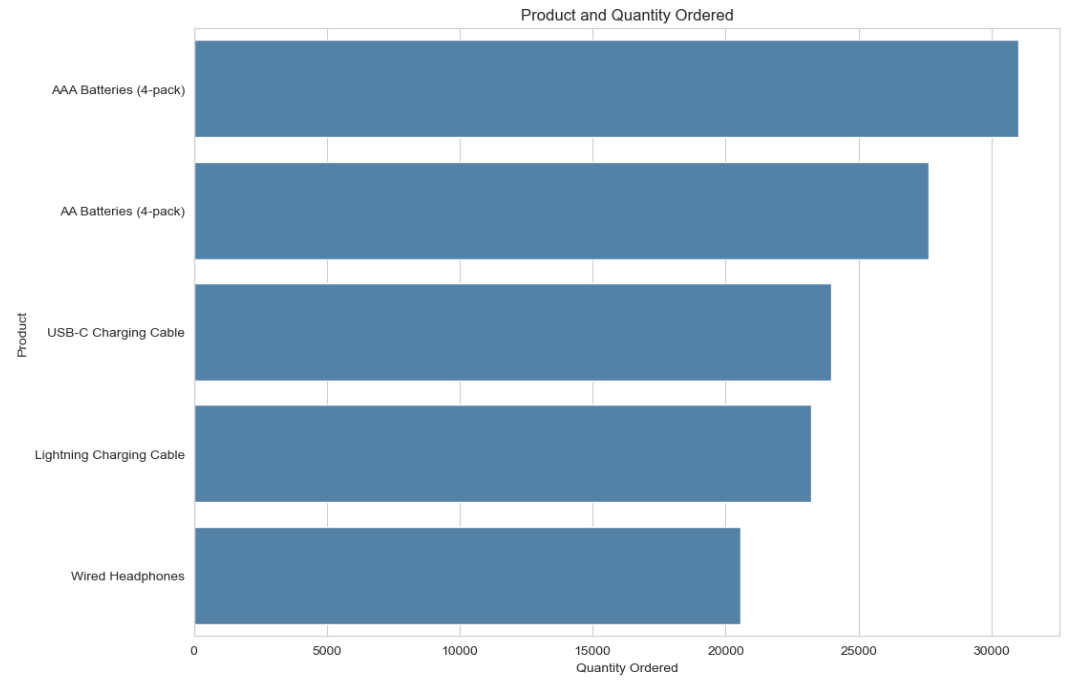
The best month for sales was December

Total sales in December amounted to $ 4619297

## In which city was the largest number of sales recorded?

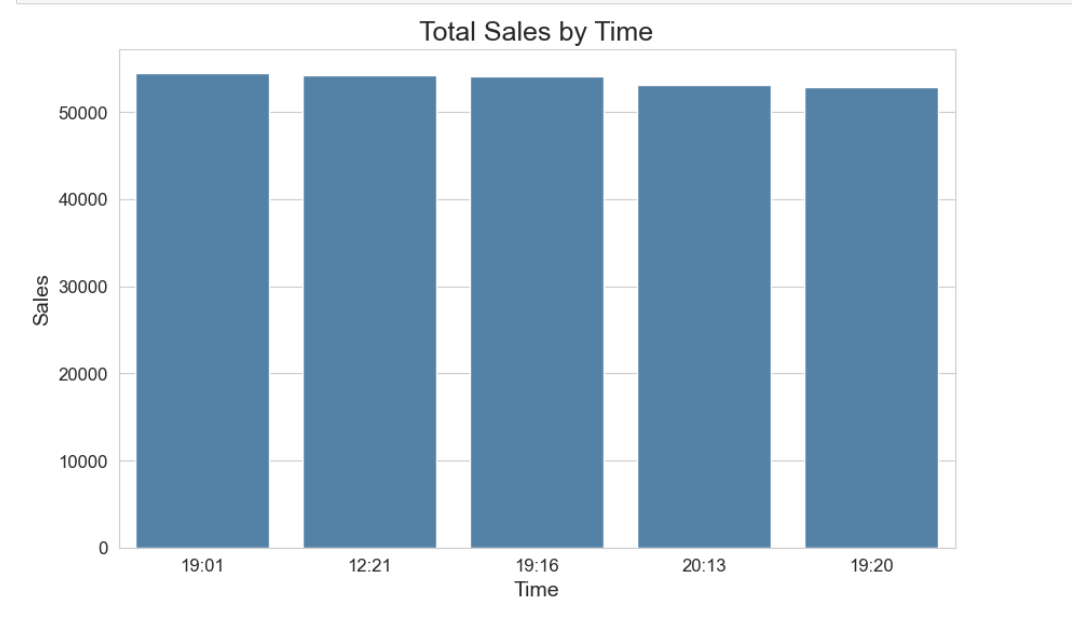


In San Francisco, the highest level of sales is about 8262204 dollars.

What product sold the most?   


31017.0 number of AAA batteries (4 pieces) sold during the year. It sells to the maximum because it is the cheapest product.

## The most appropriate time to display ads to increase the likelihood of buyers buying goods?



Total sales are maximum at these time stamps:  
19:01

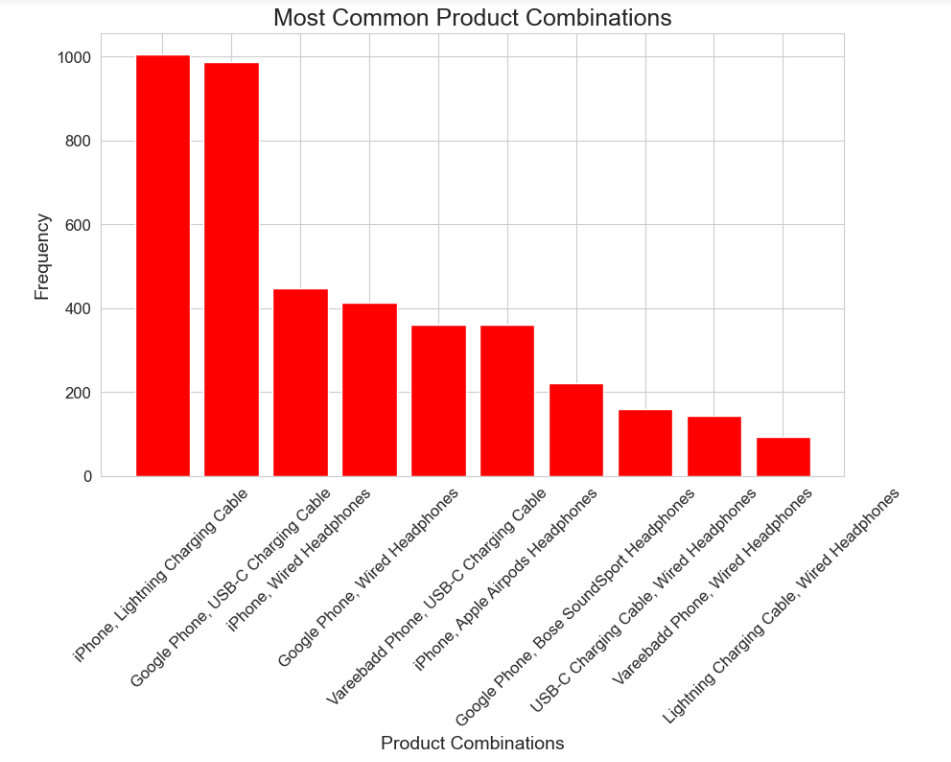
12:21

19:16

20:13

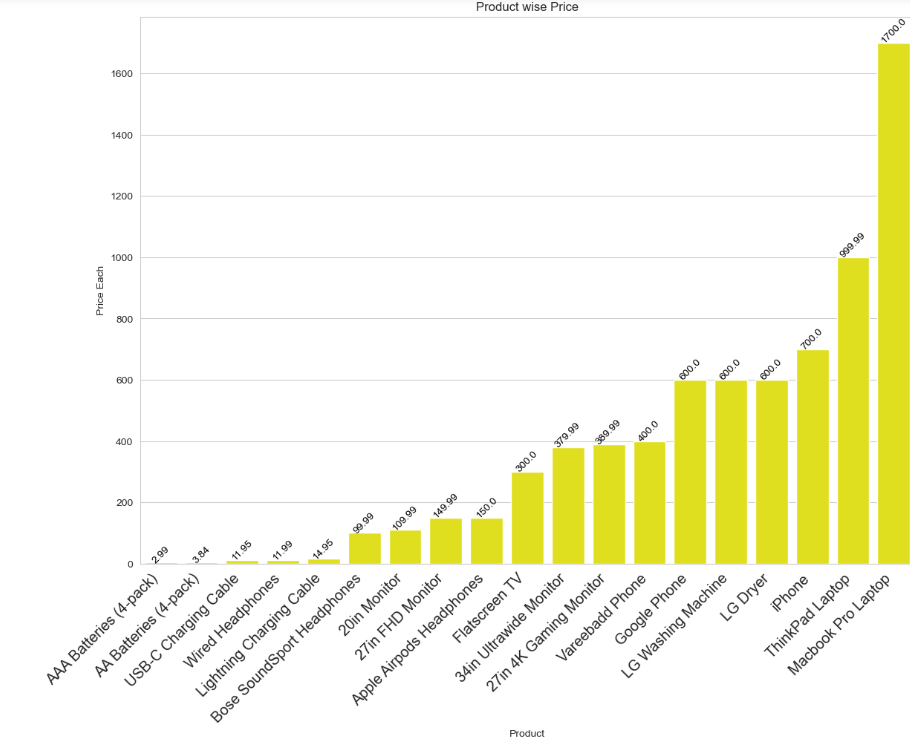
19:20

## **What products are most often sold together?**



Most often, an iPhone and a Lightning charging cable are sold together - 1005 transactions

## Product wise price



The visualization above shows that the price of AAA batteries (4 pieces) is the lowest ($2.99). Therefore, this is the product that has been sold the most.

# Machine learn technique – Classification model

I looked at the available methods and realized that only classification can be used with this database.

Regression analysis: Judging by the given columns, there is no suitable target variable for regression analysis

Clustering: The provided columns do not explicitly suggest clustering as a suitable method, as they do not mention their inherent concept of grouping or similarity.

Associative Rule Mining algorithms: These columns do not directly indicate a transactional nature and do not contain elements that can be linked for rule mining. Based on the columns given, classification seems to be the most appropriate method, since you can predict or classify a "Product" based on other available variables.

## Classification model

During this analysis, the classification task was performed using the decision tree algorithm. The goal was to predict the target variable based on the provided features or predictors. The target variable selected for classification was the "Product" column, which represents the specific product being purchased.

The features or predictors used in the classification task included various columns from the dataset, such as "Order ID", "Quantity Ordered", "Price of each" and "Order Date". These functions provided valuable information about sales transactions, product quantities, prices, and order dates.

The decision tree algorithm was chosen for the classification task because of its ability to handle both categorical and numerical features. In addition, decision trees are interpretable and can capture complex relationships between objects and the target variable.

The steps taken to train the model, evaluate, and validate included splitting the dataset into training and test sets. The training set was used to build a decision tree model, while the test set was used to evaluate its performance. The decision tree model was trained using a training set and then applied to a test set to make predictions. To assess the performance of the model, accuracy, precision, responsiveness and F1 scores were calculated.

The classification model allowed us to obtain the following performance indicators:

Accuracy: 0.988 (98.8%)

Accuracy: 0.99 (99%)

Review: 0.99 (99%)

F1 score: 0.99 (99%)

These indicators indicate that the decision tree model has performed exceptionally well in product classification. High accuracy, precision, responsiveness and F1 scores demonstrate the ability of the model to correctly predict the product category based on the functions provided.

From the classification results, it can be observed that the decision tree algorithm effectively captures patterns and relationships in the data, providing an accurate classification of products. The performance of the model suggests that the selected features are informative and make a significant contribution to solving the forecasting problem.

In conclusion, it should be noted that the decision tree classification model has demonstrated high efficiency in predicting the product category based on the characteristics provided. High accuracy indicators indicate the effectiveness of the model in the correct classification of products. These results highlight the potential of the decision tree algorithm for product classification problems and provide valuable information for making business decisions in the retail sector.

Conclusion:  
In conclusion, it should be noted that the analysis of the K mart database data allowed us to obtain valuable information and answers to the business questions posed. The main conclusions obtained as a result of the research data analysis (EDA) and the classification model are summarized as follows:

Peak sales: In December, the highest level of sales was recorded with total revenue of $4,619,297. This indicates a potential seasonal trend or an increase in consumer spending during the holiday season.

The city with the highest sales: San Francisco became the city with the most sales, reaching approximately $8,262,204. This underscores the importance of this location for K mart's business and suggests potential opportunities for further expansion or targeted marketing strategies.

Best-selling product: AAA batteries (4 pieces) became the best-selling product, a total of 31,017 units were sold during the year. The popularity of this product can be explained by its affordability, as it is the cheapest among the available options.

Optimal time for advertising: The analysis revealed specific timestamps when the total sales volume was maximum, for example 19:01, 12:21, 19:16, 20:13, and 19:20. These time periods can serve as a guideline for determining the most appropriate time for displaying ads, increasing the likelihood of attracting potential buyers.

Frequently sold products together: The combination of an iPhone and a Lightning charging cable most often appeared in transactions - in 1005 cases. This suggests potential cross-selling opportunities when targeted promotions or pooling strategies can be used to increase sales.

Product price analysis: Visualization showed that AAA batteries (4 pieces) had the lowest price - $ 2.99. This availability probably contributed to its high sales volume, making it the best-selling product overall.

The classification model gave impressive performance indicators: accuracy was 98.8%, error-free - 99%, responsiveness - 99% and F1 score - 99%. These indicators indicate the ability of the model to accurately classify and predict the target variable, providing valuable information for future decision-making.

The results of this analysis are of great importance for the K mart database and related areas. The identified sales models, the best-selling products and the optimal time for displaying ads can serve as a guide in making strategic decisions, inventory management and marketing campaigns. In addition, the performance of the classification model demonstrates its potential to improve predictive analytics and customer segmentation.

# Recommendation

To further improve the analysis, it is recommended to conduct more in-depth studies of customer behavior, market trends and competitor analysis. In addition, the integration of external data sources, such as demographic information or advertising campaigns, can provide a more complete understanding of the factors influencing sales and customer preferences.

Overall, this data analysis has shed light on the most important insights in the K mart database, providing stakeholders with useful information to drive business growth, increase customer satisfaction, and make informed decisions in dynamic retail.

# References:

1. “Predict Customer Churn” data analysis file that we received during the classes
2. [GitHub - jsc1535/K-Mart-Data-Analysis: K-mart sales strategy for the year 2020 based on the insights from the sales data in 2019.](https://github.com/jsc1535/K-Mart-Data-Analysis)

Abstract:

This data analysis report was made for exploring the K-mart sales in 2019 database to find valuable insights and answer to several business questions. The report includes a comprehensive exploratory data analysis of sales data, products, and customer behavior. Using different techniques, the report provides with meaningful information that can help to K-mart team to develop the sales strategy for the next years.

The dataset contains several key columns, including “Order ID”, “Product”, “Quantity Ordered”, “Price Each”, “Order Date”, “Purchase Address”, “month”. These columns contain important information about sales transactions, product information and customer orders.

The report begins with database introducing and describing the purpose of the report. It provides the data structure.

Introduction:

Classification Report

The following report provides an overview of the classification results obtained using a Decision Tree Classifier on the given dataset. The objective was to classify different products based on the provided features.

Accuracy Score: 0.9884646410325356

The model achieved an impressive accuracy score of 98.8%, indicating its ability to correctly classify the products in the test set.

Precision, Recall, and F1-Score by Class

The classification report presents the precision, recall, and F1-score for each class in the target variable. These metrics provide insights into the model's performance for individual product categories.

The majority of the classes achieved perfect precision, recall, and F1-scores, indicating that the model accurately classified instances belonging to these classes. These classes include "20in Monitor," "27in 4K Gaming Monitor," "27in FHD Monitor," "34in Ultrawide Monitor," "AA Batteries (4-pack)," "AAA Batteries (4-pack)," "Apple Airpods Headphones," "Bose SoundSport Headphones," "Flatscreen TV," "Lightning Charging Cable," "Macbook Pro Laptop," "ThinkPad Laptop," "USB-C Charging Cable," "Vareebadd Phone," "Wired Headphones," and "iPhone."

However, the model's performance was comparatively lower for certain classes. For example, the "Google Phone" class achieved a precision of 0.79, suggesting that the model correctly predicted 79% of instances for this class. Similarly, classes such as "LG Dryer" and "LG Washing Machine" had lower precision, recall, and F1-scores, indicating that the model struggled to accurately classify instances for these classes.

Overall Evaluation

The decision tree classifier demonstrated excellent performance in classifying the majority of product categories, with an accuracy of 98.8%. The model's ability to accurately predict most classes suggests its effectiveness in capturing the patterns and relationships between the input features and the product labels.

However, it is important to further investigate and address the lower performance observed for certain classes, such as "Google Phone," "LG Dryer," and "LG Washing Machine." Additional data or feature engineering might be necessary to improve the model's performance for these specific classes.

The classification results obtained from this analysis provide valuable insights into the potential of using machine learning techniques to classify products based on the provided features. Further refinement and optimization of the model could potentially enhance its classification performance and facilitate more accurate product categorization.