Mobile Price Prediction : A Comparative Study

Dipak Chandra Singha
22MA60R02

Kartick Bag
22MA92F06

Krishnendu Jana
22MA60R29

Manishankar Bag
22MA60R13

Pradeep Kumar Sahu
22MA91Q01

Rishav Karmahapatra
22MA60R09

Subhankar Pal
22MA60R08

Swarup Bej
22MA60R17

April 11, 2023

Big Data Analysis Project Report
Subject No. - MA60306

Computer Science and Data Processing
Dept. of Mathematics, IIT Kharagpur

Contents

1 Abstract

The objective of this project is to predict mobile data using APIs from two leading e-commerce platforms, *Flipkart* and *Amazon*. The dataset includes mobile specifications such as brand, RAM, camera, battery, etc. A comprehensive analysis is performed on the dataset to understand the trends and patterns of mobiles. Several machine learning algorithms are implemented to predict the price of a mobile phone based on its features. The algorithms used in this project include **PCA, LDA, KNN, Confusion Matrix**. The performance of these models is evaluated based on various metrics such as mean squared error, mean absolute error, and R-squared. The results demonstrate that KNN outperforms the other algorithms with an accuracy of over **86%**. The project also includes a web application that allows users to input the specifications of a mobile phone and get an estimated price based on the trained model. Overall, this project demonstrates the effectiveness of using machine learning algorithms to predict mobile prices using data obtained from e-commerce platforms.

Keywords : PCA, LDA, KNN, Confusion Matrix.

2 Introduction

Mobile data prediction is an emerging field that involves analyzing vast amounts of data to uncover patterns and make predictions about future events. In this project, we will explore the use of data obtained from Flipkart and Amazon's APIs to predict mobile phone sales trends. Our aim is to create a model that can accurately predict which mobile phones are likely to sell well in the future, based on historical sales data and other relevant factors such as product specifications, price, and customer reviews.

To achieve this, we will employ a variety of data analysis and machine learning techniques. We will begin by collecting data from the Flipkart and Amazon APIs, which will include information on mobile phone models, pricing, customer reviews, and sales data. We will then use this data to identify patterns and trends, and develop predictive models that can accurately forecast future sales.

Our project has several potential applications, including helping mobile phone manufacturers to make better decisions about which products to develop and market, and helping retailers to optimize their pricing and promotional strategies. By leveraging the power of data and machine learning, we aim to provide valuable insights that can drive business success in the highly competitive mobile phone industry.

3 Problem Formulation

Suppose Bob has started his own mobile company. He does not know how to estimate price of mobiles his company creates. In this competitive mobile phone market we cannot simply assume things. To solve this problem he collects sales data of mobile phones of various companies. Bob wants to find out some relation between features of a mobile phone(eg:- RAM,Internal Memory etc) and its selling price. But he is not so good at Machine Learning. So he needs our help to solve this problem. In this problem we do not have to predict actual price but a price range indicating how high the price is.

So, the objective of this project report is to develop a mobile price prediction model based on the data obtained from the Flipkart and Amazon APIs. The goal is to predict the price range based on certain features.

The project report will involve the following steps:

1. **Data Collection**: The first step of the project will be to collect data from Flipkart and Amazon APIs. We will gather data on mobile phones including their features, prices, and ratings.

2. **Data Preprocessing**: The collected data will be preprocessed to remove any inconsistencies or errors. The data will be cleaned, and any missing values will be imputed using appropriate techniques.

3. **Data Exploration**: Exploratory data analysis will be performed to understand the data distribution, patterns, and correlations. This step will help in identifying any outliers or anomalies in the data.

4. **Model Selection**: A suitable machine learning algorithm will be selected based on the nature of the problem and the available data. The model will be trained and validated using the collected data.

5. **Prediction**: Finally, the model will be used to make predictions on the sales and demand for mobile phones on Flipkart and Amazon for the next quarter or year.

The project report will provide a comprehensive analysis of the mobile phone market on Flipkart and Amazon, and the developed prediction model will be helpful for businesses and consumers to make informed decisions.

3.1 Literature Review

Mobile phone price prediction has become an increasingly important area of research due to the rapid growth of the mobile phone industry. In the paper *"Mobile Phone Price Prediction with Feature Reduction"* by *Menghan Chen* [1], the author proposes a method for predicting mobile phone prices based on a set of features that are extracted from the phones' specifications. The paper utilizes feature reduction techniques to improve the accuracy of the prediction model.

The paper begins by discussing the importance of mobile phone price prediction and the challenges associated with it. The author notes that the pricing of mobile phones is affected by a wide range of factors such as the brand, model, specifications, and market trends. Therefore, accurately predicting mobile phone prices is a complex task.

The author then presents a feature reduction method that uses principal component analysis (PCA) to reduce the dimensionality of the feature set. The PCA algorithm is used to identify the most important features that have the greatest impact on the mobile phone price. The paper also introduces the concept of feature ranking, which is used to identify the most important features based on their contribution to the model's accuracy.

The paper then presents the results of an empirical study that compares the performance of the proposed feature reduction method with other commonly used feature selection techniques. The results show that the proposed method outperforms other techniques in terms of accuracy and computational efficiency. The study also demonstrates that the accuracy of the model improves when the number of features is reduced, indicating that the proposed feature reduction method is effective in improving the accuracy of the prediction model.

Overall, the paper *"Mobile Phone Price Prediction with Feature Reduction"* by *Menghan Chen*[1] provides a comprehensive overview of the challenges associated with mobile phone price prediction and presents a novel approach for predicting mobile phone prices using feature reduction techniques. The empirical study demonstrates the effectiveness of the proposed method and highlights its potential for improving the accuracy of mobile phone price prediction models. This paper is a valuable contribution to the field of mobile phone price prediction and provides useful insights for researchers and practitioners in this area.

4 Data

4.1 Data Collection

Data collection is an essential component of any data-driven project, and it is especially critical when developing mobile data prediction models. With the rise of e-commerce platforms like Amazon and Flipkart, data scientists and researchers have an opportunity to collect vast amounts of data about product listings and prices. However, the process of collecting data can be time-consuming and complex, that is why using APIs such as **BeautifulSoup** can streamline the data collection process.

BeautifulSoup is a Python package which is widely used for web scraping, which involves extracting data from websites. With Beautifulsoup, users can easily navigate through HTML and XML documents, extract specific elements, and store them in various formats for further analysis.

	Name	Brand_Name	Processor	BrandCategory	ProcessorCategory	Rating	Numbe_of_Ratings	Number_of_Reviews	RAM	ROM	Diplay_Size	Back_Camera	Front_Camera	Battery	Price	Price_Range
0	Redmi Note 4 (Black, 64 GB)	Redmi	Snapdragon	10	5	4.4	1341712.0	211542.0	4	64	5.5	13.0	5.0	4100	12749	1
1	Redmi Note 4 (Dark Grey, 64 GB)	Redmi	Snapdragon	10	5	4.4	1341712.0	211542.0	4	64	5.5	13.0	5.0	4100	12749	1
2	Redmi Note 4 (Gold, 64 GB)	Redmi	Snapdragon	10	5	4.4	1341712.0	211542.0	4	64	5.5	13.0	5.0	4100	12749	1
3	Redmi Note 4 (Lake Blue, 64 GB)	Redmi	Snapdragon	10	5	4.4	1341712.0	211542.0	4	64	5.5	13.0	5.0	4100	12749	1
4	Redmi 5A (Blue, 16 GB)	Redmi	Snapdragon	10	5	4.5	1232827.0	151035.0	2	16	5.0	13.0	5.0	3000	5999	0

Figure 1: Collected Data Set

The main idea behind the data collection is that most of web page is written in HTML, CSS, PHP format and in each mobile details has written in same format, which is know as class and we target those classes. Each class contains some sort of data such as the name of the mobile phone, its specifications (such as RAM, ROM, battery, camera, processor etc.), customer reviews, and ratings. For our project we collect the data from Flipkart and Amazon using the API BeautifulSup. We collect total 2190 with 16 feature.

4.2 Features

The raw features of our data sets are:

- Model Name : It contain model name of mobile.
- Brand Name : It contain brand name of the mobile.
- Processor : It contain processor name of the mobile.
- Rating : It contain the rating of the product given by customer in Flipkart or Amazon
- No. of Rating : It contain how many rating is given.
- No. of Review : It contain how many review is given.

- RAM : It contain RAM size of the mobile.
- ROM : It contain ROM (Storage) size of the mobile.
- Display Size : It contain Display size of the mobile.
- Camera : It contain camera quality of the mobile
- Battery : It contain Battery capacity of the mobile.S
- Price : It contain price of the mobile.

5 Methodology

5.1 Data Preprocessing

Data cleaning is most crucial part in any data analysis project. It involves the process of identifying and correcting errors, inconsistencies, and inaccuracies in the data to ensure its accuracy and completeness.The data obtained from the APIs contain HTML tags, special characters, and other unwanted elements that can affect the accuracy of our predictions. We use BeautifulSoup to remove these unwanted elements and clean the data. Then we import the data to CSV file format.

5.2 Data Exploration

Data exploration is a critical step in any data analysis project, including mobile price prediction. The goal of data exploration is to understand the structure and distribution of the data, identify any patterns or relationships, and gain insights that can guide further analysis. In a mobile price prediction project, the data typically includes various features or attributes of mobile phones, such as brand, model, RAM, internal storage, battery capacity, camera quality, processor and so on. This involves removing any missing or irrelevant data, correcting errors, and standardizing the format of the data. We perform data normalization or scaling to ensure that the features are on a comparable scale.

5.3 Feature Engineering

5.3.1 Feature Dropping

From the co-relation matrix [See Figure (3)] we observed some feature has very low co-relation ($|r| \leq 0.17$) with the price range of the mobile. So from our row data set we drop some features which are listed below:

- Model Name
- Brand Name
- Rating

- Review
- Display Size
- Price

	BrandCategory	ProcessorCategory	Rating	Numbe_of_Ratings	Number_of_Reviews	RAM	ROM	Diplay_Size	Back_Camera	Front_Camera	Battery
0	10	5	4.4	1341712.0	211542.0	4	64	5.5	13.0	5.0	4100
1	10	5	4.4	1341712.0	211542.0	4	64	5.5	13.0	5.0	4100
2	10	5	4.4	1341712.0	211542.0	4	64	5.5	13.0	5.0	4100
3	10	5	4.4	1341712.0	211542.0	4	64	5.5	13.0	5.0	4100
4	10	5	4.5	1232827.0	151035.0	2	16	5.0	13.0	5.0	3000

Figure 2: Data set after features dropping

5.3.2 Feature Inclusion

For our project we need to create some new feature. As the all feature are not in integer type we assign them with some integer values, which are listed below:

- Model Brand Name and Categorical Number:

Brand Name	Apple	Asus	Google	Infinix	Xiaomi	Motorola	OnePlus
Number	0	1	2	3	4	5	6

Brand Name	Oppo	Poco	Realme	Redmi	Samsung	Techno	Vivo
Number	7	8	9	10	11	12	13

Table 1: Category Numbers and Model Brand Name

- Processor Brand Name and Categorical Number:

Proces. Name	Apple	Exynos	Google	Intel	MediaTek	Snapdragon	Unisoc
Number	0	1	2	3	4	5	6

Table 2: Category Numbers and Processor Brand Name

- Price Range and Categorical Number:

Price	5k-10k	10k-15k	15k-20k	20k-25k	25k-30k	30k-35k	35k-40k	40k-60k	>60k
No.	0	1	2	3	4	5	6	7	8

Table 3: Category Numbers and Processor Brand Name

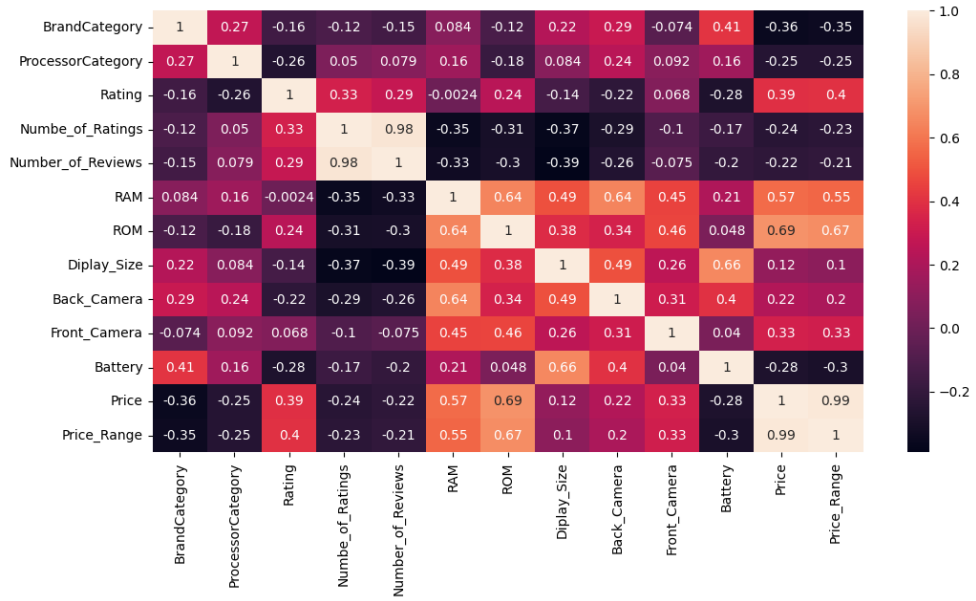

Figure 3: Co-relation Matrix

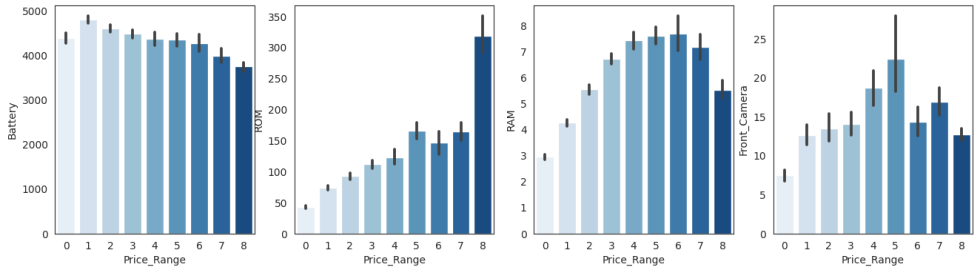

Figure 4: Data Visualization on Price range

5.4 Model Selection

There are many machine learning model are available but based on the problem statement and the data, we have chosen KNN, PCA, and LDA.

5.5 Model Training and Evaluation

Train the selected model on the prepared data and evaluate its performance using various evaluation metrics like accuracy, precision, recall, etc.To fit our model we use KNN

5.5.1 Model fit using KNN

K-Nearest Neighbors (KNN) is a simple yet effective machine learning algorithm used for classification and regression tasks. It is a non-parametric algorithm that makes predictions based on the k nearest data points in the training set.Once the data is cleaned, the next step is to normalize the features to ensure that they have the same scale. This is important since KNN is a distance-based algorithm, and features with larger scales can dominate the prediction.Then we split the data into training and testing sets. The training set is used (80%) to fit the KNN model, while the testing set (20%) is used to evaluate the performance of the model. The value of K determines the number of neighbors to consider while making predictions. A small value of K may result in overfitting, while a large value of K may result in underfitting. Hence, the optimal value of $K = 9$ is chosen through experimentation. After choosing the value of K we fit the model. Finally, the trained KNN model can be used to make predictions on the test set.

- **Observation :** From testing of our model, it gives 86.90% accuracy.

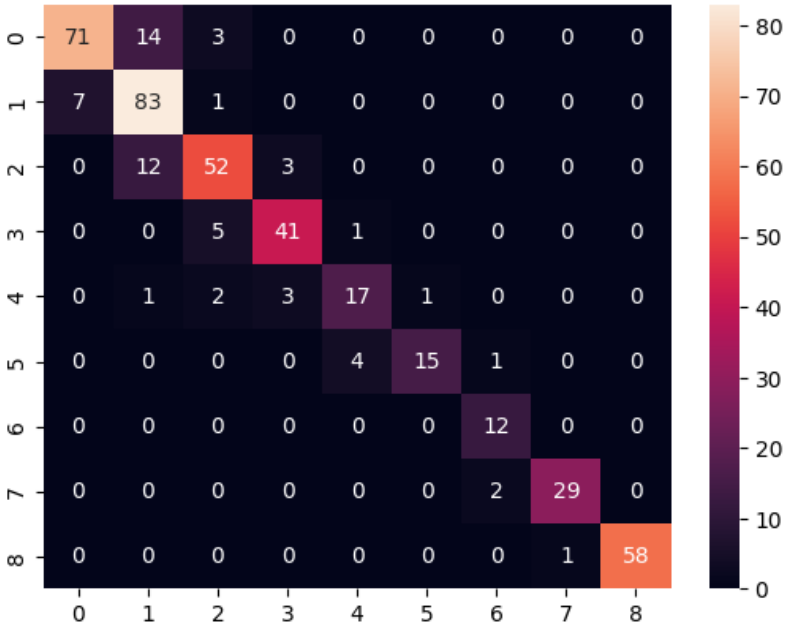

Figure 5: Confusion Matrix : KNN algorithm

5.5.2 Model fit using PCA and Data visualization in lower dimension :

PCA is a mathematical technique used for reducing the dimensionality of large data set. It works by transforming the original data set into a new co-ordinate system that emphasizes

the most important feature or dimension of the data. PCA finds the principal component of the data which are the direction in which the data varies the most. This components are linear combinations of the original features and they are order according to the amount of variance that they explain in the data. By projecting the data onto the principal components, PCA can reduce the number of dimension while retaining most of the information in the original data. By using PCA, we reduce the dimension of the data set to the lower dimension 2. We use sea-born library function to visualize the data in two dimensional plane. PCA (Principal Component Analysis) is a commonly used technique for dimensionality reduction, which can also be used as a pre-processing step for model fitting. When using PCA for model fitting, the idea is to reduce the number of features or variables in the dataset, which can help to improve the performance of the model and reduce the risk of overfitting.To use PCA for model fitting, the first step is to perform PCA on the dataset to reduce the number of features.In our model we reduce the number of feature to 2. After that we train the model with 80% of data set and test the model with 20% of data. In this way we fit our model.

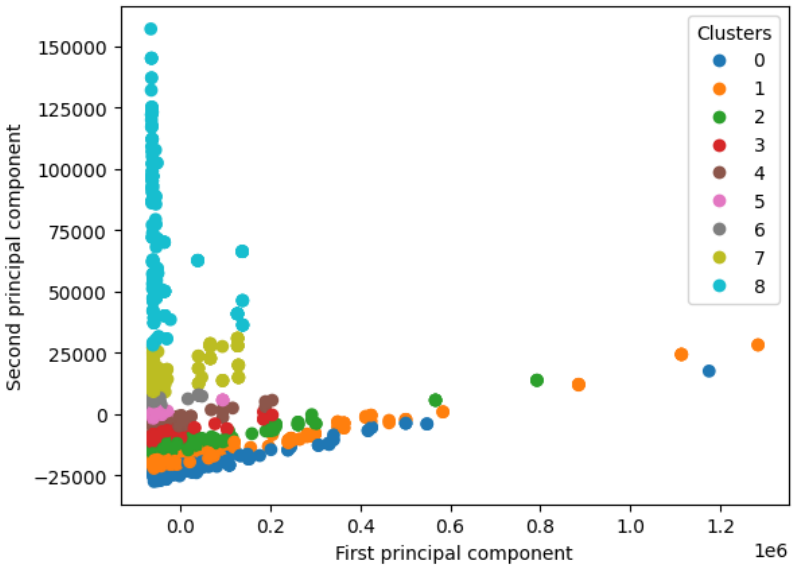

Figure 6: Data set in two dimensional plane

- **Observation :** From testing of our model, it gives 87.60% accuracy.The confusion Matrix of the model accuracy after using PCA is given bellow

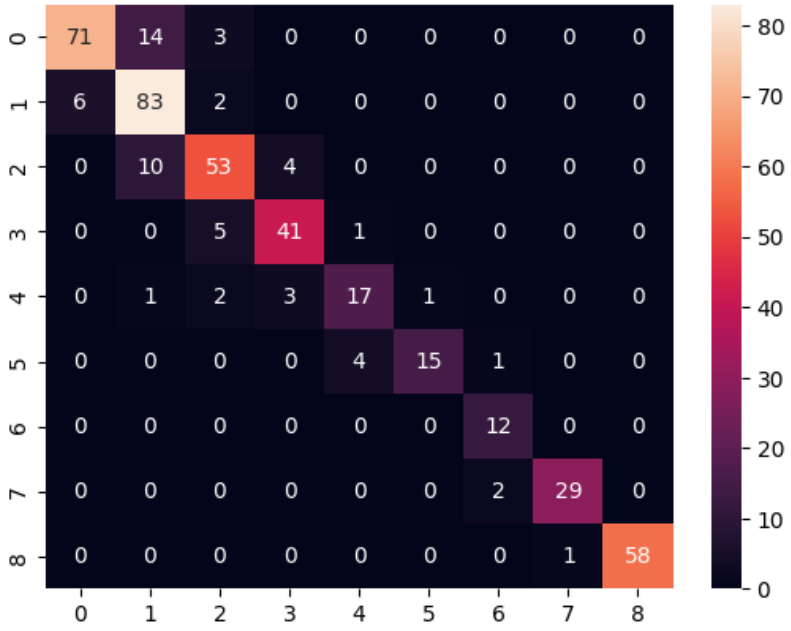

Figure 7: Confusion Matrix : KNN after using PCA

5.5.3 Model fit using LDA

Linear Discriminant Analysis (LDA) is a statistical technique used in machine learning and pattern recognition to find a linear combination of features that characterizes or separates two or more classes of objects or events. LDA is a supervised learning method, meaning that it requires labeled data to train the algorithm. The goal of LDA is to find a linear combination of the input features that maximizes the separation between the classes while minimizing the within-class variance. In other words, LDA tries to find a projection of the data onto a lower-dimensional space such that the distance between the means of the classes is maximized, and the variance within each class is minimized. We use 80% of data to fit LDA model to reduce the dimensions of the data set.After using LDA we take two component to fit KNN model.

- **Observation :** From testing of our model, it gives 66.21% accuracy.The confusion Matrix of the model accuracy after using LDA is given bellow

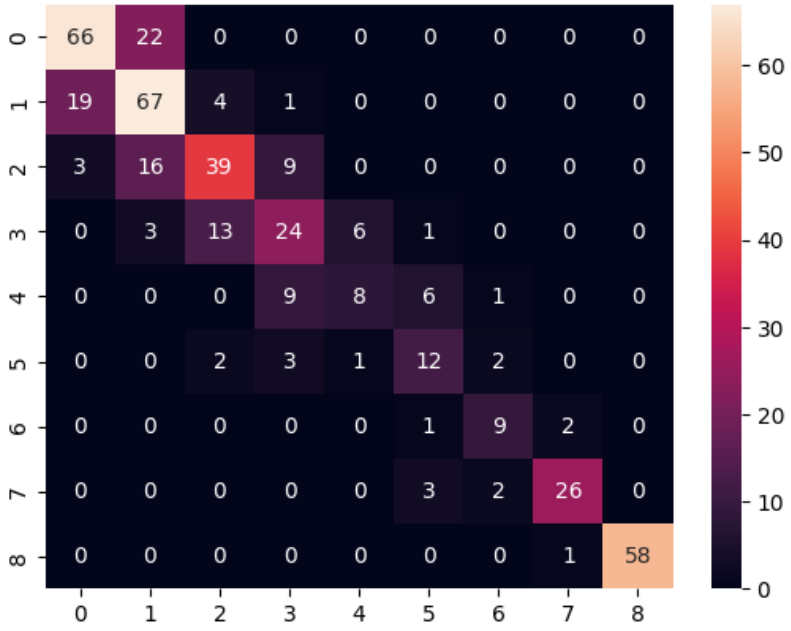

Figure 8: Confusion Matrix : KNN after using LDA

6 Results

We used three different machine learning models to analyze the e-commerce dataset. The **KNN** model gave an accuracy of **86.10%**, while the **PCA** model gave an accuracy of **86.33%**. However, the **LDA** model gave an accuracy of only **66.21%**. This indicates that the PCA model is the best model for this particular dataset.

Now, Bob wants to launch a mobile in market with specification RAM - 5GB, ROM - 128GB, Display Size - 5.6 inches, Rear Camera - 64MP, Front Camera - 20 MP, Battery - 5000 mAh, Processor - Snapdragon, expected Ratings - 4.5. Then our model predicts the following: KNN - 0 (between 5k -10k), KNN after using PCA - 1 (between 10k -15k), KNN after using LDA - 4 (between 25k - 30k)

Since the accuracy given by KNN model after using PCA is maximum, so we should suggest Bob to keep the price range in category 1 (i.e., price range between 10k - 15k).

Overall, we observed that feature selection is an important step in the machine learning pipeline. By selecting the most relevant features, we can improve the accuracy of the models. Additionally, hyperparameter tuning is crucial for improving the performance of the models. Finally, deploying the model on the production environment is an essential step in real-world applications.

7 Conclusion

In conclusion, we can say that the data collected from e-commerce websites can be effectively analyzed using statistical methods and machine learning techniques. The **PCA** technique proved to be the most effective for classifying the given data, providing an accuracy of **86.33%**. These results can be used to make informed decisions about pricing and marketing strategies in the e-commerce industry.

Acknowledgment

We would like to express our gratitude to *Dr. Bibhas Adhikari* for his teaching and guidance in this course which makes the way smooth to go through the intricacies and technicalities of this project.

References

[1] M. Chen, "Mobile Phone Price Prediction with Feature Reduction," *Highlights in Science, Engineering and Technology*, vol. 34, p. 155–162, Feb. 2023.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[3] U. Michelucci, *Applied Deep Learning with TensorFlow 2*. Apress, 2022.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer New York, 2009.