

Effectiveness Of Predictive Analysis For Pricing in E-Commerce:A Machine Learning Approach



Sandeep Gopu

Dublin Business School

This dissertation is submitted for the degree of
Master of Science in Data Analytics

January 2024

Declaration

I, Sandeep Gopu, declare that this work is my original work and has not been submitted to any institution or university for the award of degree. Additionally, I have accurately cited all literature and sources used in this research, and this research will be treated in strict confidence in accordance with the Academic Honesty Policy of Dublin Business School.

Sandeep Gopu
January 2024

Acknowledgements

I would like to thank my professor and Research Supervisor, Paul Laird, for his patient guidance from the beginning, for providing valuable and constructive suggestions during the planning and development of the study, and for generously giving me his time. I would also grateful to my professor for his advice and support in data selection, data preprocessing, and for providing samples and resources in the area of price prediction analysis. A Sincere thank you is also extended to my family and friends for their constant love and support throughout this transformative journey.

Abstract

This thesis explores how mobile phone price prediction models develop in the dynamic e-commerce world. By using web scraping and advanced machine learning techniques, this project constructed a model demonstrating high accuracy in cost estimation. This power lets businesses set prices based on data, and helps getting the most profit by changing them in real-time based on demand, seasonability, and what the competition is doing. The model also helps set personal prices by figuring out different types of customers and how mindful they are of price. Although, by always researching and trying out new cutting-edge methods like deep learning, there's more chance to make the model even better and adapt to market trends. In the end, the thesis explains how priceless mobile price prediction models are for helping businesses stand their ground in tough e-commerce space and helps to achieve sustainable growth.

Table of contents

List of figures	vii
List of tables	viii
1 Introduction	1
1.1 Predictive analysis for predicting mobile phone prices	1
1.2 Background	1
1.3 Motivation	2
1.4 Research Objectives	3
1.4.1 Market Intelligence via Data Scraping	3
1.4.2 Data Quality Assurance and Cleaning	3
1.4.3 Utilising exploratory data analysis(EDA) to make strategic decisions	3
1.4.4 Predictive Pricing Model for Optimisation	3
1.4.5 Finding The Main Elements Affecting prices	3
1.5 Research Questions	4
2 State of the Art	6
2.1 Recent Developments	6
2.2 Related work	7
3 Problem Statement	10
4 Design	12
4.1 The Gathering of Data	12
4.2 Data Cleaning	13
4.3 Modelling Data	17
5 Implementation	22

Table of contents	vi
<hr/>	
6 Evaluation	32
6.1 Mean Square Error:	32
6.2 Root Mean Square Error:	32
7 Conclusions	35
References	37
Appendix A Code Snippets and Implementation	39
Appendix B Data Sources and ethical Considerations	45

List of figures

4.1	RAM values	14
4.2	Multiple columns	15
4.3	Final dataset	17
5.1	Heat map	22
5.2	Popular Brands	24
5.3	Relationship between RAM and Storage	25
5.4	Screen Size for top 10 Brands	26
5.5	Pie chart Showing the Distribution percentage of Operating system	27
5.6	Price Distribution	28
5.7	Bar Chart for Average Price with 2G,3G,4G,5G,NFC	29
5.8	Feature Importance	30
A.1	Data Scraping from eBay	39
A.2	Extracting important features from the page	40
A.3	Linear Regression	40
A.4	Decision Tree	41
A.5	Random forest	41
A.6	SVM Regression	42
A.7	XGBoost	42
A.8	K Nearest Neighbors	43

List of tables

6.1 MSE and RMSE values For each Algorithm 34

Chapter 1

Introduction

1.1 Predictive analysis for predicting mobile phone prices

To stay competitive in e-commerce's rapidly evolving landscape, businesses must predict and optimise prices. This research on mobile phones pricing takes a thorough approach to price prediction. By combining machine learning and data scraping, it enables adaptability towards the changing market demands and pricing. The analysis is based on data extracted from eBay listing which has useful data on brands, features, and pricing patterns. By using web scraping, we systematically collected these data for building a rich dataset for the research.

The project explores machine learning models like Linear Regression, Decision Trees, Random Forests, SVM Regression, XGBoost and K Nearest Neighbours. These algorithms identify pricing variables and trends and help in gaining an understanding of the variables that are affecting mobile prices. This project explains how businesses interpret patterns and insights to improve strategic pricing.

The main goal is to build an analytical tool which can be used for more than simple price prediction. It should explore how predictions affect pricing strategy, customer experience, and decision-making. The machine learning algorithms are used to gain the strategic advantage by integrating predictive analytics. By doing this, we are trying to empower the businesses with a more knowledgeable, flexible, and customer-focused mobile pricing approach in this evolving e-commerce market.

1.2 Background

In the evolving e-commerce world, long-term success needs the understanding for changing market conditions. The mobile phones widespread use and the advancement in technology

every other day, businesses encounter a unique business challenge. Predicting and optimising mobile prices becomes essential due to the ever-evolving landscape of customer preferences, technological advancement catering, and competitive market space.

For companies navigating internet shopping's choppy waters, predictive analysis is a key compass. It highlights how predictive analytics can revolutionise decision-making by uncovering patterns, forecasting market trends, and enabling proactive choices. This research spotlights predictive analysis's worth for E-commerce players keen to stay competitive while focusing on mobile pricing intricacies.

1.3 Motivation

This project is motivated by a large goal to use predictive analysis to transform the pricing strategy in the e-commerce in thee-commerce industry. The dynamic realm of online retail presents firms with the problem of strategically determining products prices based on customer behavior, market trends, and competitive factors as the scenarios change very frequently. The main goal is to provide e-commerce businesses with a predictive tool that goes beyond traditional methods so they can remain flexible, make wise decision, and become more competitive overall.

The aim to improve pricing accuracy is what motivates the use to predictive analysis in e-commerce. Conventional approaches frequently fail to capture the subtle features that affect pricing decision. Businesses can use predictive models to analyse large dataset and finds patterns that reveal relationships between quantities of products, customer behavior and best prices. The main goal is enable e-commerce companies to adopt proactive ,data-driven pricing strategies that are responsive to changing market conditions in real time, rather than relying solely on reactive approaches.

Predictive analytics may help businesses build lasting relationships with their customers that will promote loyalty and long term growth ,going beyond simple transactional exchanges. The motive captures the idea that predictive analysis may be used as a pillar to help navigate the intricacies of e-commerce, ultimately leading to success through strategic flexibility and data driven insights. Based on these insights, the efficient pricing technique will also help the business in taking an advantage over the competitors. This will help the business in offering competitive pricing while being profitable as well.

1.4 Research Objectives

1.4.1 Market Intelligence via Data Scraping

In the first stage, sellers receive useful insights into the mobile phone market by systematically extracting data from eBay. Businesses can obtain a thorough understanding of the competitive landscape and modify their strategies by compiling data on products names, prices and specifications.

1.4.2 Data Quality Assurance and Cleaning

By carrying out a thorough cleaning procedure, the code guarantees the accuracy and dependability of the data that has been gathered. For sellers to improve the overall customer experience, set competitive prices, and make well-informed decisions, accurate and consistent information is essential.

1.4.3 Utilising exploratory data analysis(EDA) to make strategic decisions

Through the section on exploratory data analysis, businesses can find patterns, trends, hidden insights and customer preferences in the dataset. Sellers can help with inventory management and the development of targeted marketing strategies by identifying preferred features, popular brands, and pricing trends. Gaining the understanding of key features which affects the pricing plays a vital role in business development.

1.4.4 Predictive Pricing Model for Optimisation

The code makes it easier to predict mobile phone prices by utilising machine learning models like Linear Regression, Decision Tree Regression and others. With the help of this predictive capability, seller can maximise revenue and maintain market competitiveness by optimising their pricing strategies.

1.4.5 Finding The Main Elements Affecting prices

The features that have a major impact on mobile phone prices are revealed by the feature importance analysis. Businesses can use this information to prioritise features that customers find most appealing in their products and modify their pricing strategies accordingly.

1.5 Research Questions

Effects of Data Scaling and Outlier Removal

- What impact does the elimination of outliers have on the accuracy and robustness of predictive models?
- what role does data scaling play in enhancing the model's capacity to manage a wide range of features at different scales?

Strategies of competitive pricing

- How can companies figure out what prices to charge for their mobile phone products that are competitive?
- How do things like brand, features and consumer trends affect the best places to put a price?

Data dependability and quality

- How do the top 10 brands of mobile phones influence their price?
- what steps can be taken to efficiently clean and preprocess the data so that accurate insights are obtained?

Decision-Making focused on the client

- How can companies use exploratory data analysis(EDA) to learn about market trends and consumer preferences?
- What knowledge can be gained in order to decide on product offerings and marketing tactics that are focused on the needs of the customer?

Model of predictive pricing

- To what extent can machine learning models like Decision Tree Regression and Linear Regression predict the price of mobile phones?
- What is the distribution of operating systems(Android,ios,etc) among the listed mobile phones?

The significance of feature and customer value

- What are the main characteristics that have a big influence on the cost of mobile phones in the internet market?
- In what ways can companies give priority to these features to improve the way that customers view the value of their products?
- What role does each feature RAM,Processor,Camera,Storage and Connectivity. for example-Play in the model's overall ability to forecast outcomes?

Chapter 2

State of the Art

2.1 Recent Developments

Predictive analysis and e-commerce have been going hand in hand in recent years as companies look for novel approaches to enhance pricing policies and obtain competitive advantages in the online market. An overview of significant research, approaches and developments leading to predictive analysis in the context of e-commerce pricing optimisation may be found in the literature review that follows

For e-commerce businesses, predictive analysis has become essential to forming strategic decision making. Researchers leverage its crucial function in predicting market trends, predicting consumer behaviour and above all, optimising product pricing. As per the finding of Zhu (2021), predictive analysis facilitates enterprises to transcend conventional pricing models and adopt flexible, data-driven strategies that are adaptable to changing market conditions. The use of machine learning models for efficient pricing optimization in e-commerce is emphasised in the literature. Research on the comparative comparison of decision tree, regression models and ensemble techniques like Random forest for product price prediction can be found in the performed research. Linear regression, Decision Tree, Random forest, SVM (Support Vector Machine), XGBoost (Extreme Gradient Boosting) and KNN (K Nearest Neighbors) are among the many predictive models that are frequently used in the area and the code implementation of these models follows this pattern.

Predictive analysis's understanding of how different products' attributes affect price is essential. Researchers who emphasise the value of feature engineering and analysis in determining ideal prices include this technique for being relevant and competitive in the market. Szepannek and Aschenbruck (2019) In line with this study strategy, the code explores the significance of several features, such as RAM, Processor, camera, and connectivity. The goal is to identify the key factors that influence mobile phone price. Getting broad and varied

datasets for predictive model training now requires web scraping. The potential of online scraping to gather real-time market data for e-commerce analytics is covered by this research paper Ong *et al.* (2019). The fact that the algorithm uses web scraping to obtain data from eBay highlights how useful this method is for improving the dataset's richness.

In predictive analysis, addressing outliers Shen (2018) highlight how outlier removal effects model performance, and talk about how scaling is crucial to treating feature with different scales fairly. Outliers elimination and data scaling are included in the code in accordance with recommended procedures mentioned in these publications. Predictive analysis's application to e-commerce pricing is becoming more and more studied. Some investigated how revenue and customer satisfaction are positively impacted by predictive analytics. These results align with the code's ultimate objective of optimising mobile phone prices in an e-commerce context, highlighting the possible commercial benefit of putting predictive pricing tactics into practice.

2.2 Related work

Bar-Gill *et al.* (2023) investigate how to improve small businesses data-driven operations using an eBay field experiment. The study fills in the gaps in the body of knowledge on data-driven decision making for small businesses by concentrating on the distinct eBay ecosystem. The study approach incorporates interventions to evaluate effects on eBay business owners, offering signification perspectives to stakeholders and legislators. The study's conclusions are intended to guide the development of data-driven small business cultures, with a focus on eBay function as a platform for the growth of entrepreneurship. Fong and Waisman (2023) examines the effects of bargaining delays on eBay transactions. The study adds to the body of knowledge on online negotiation dynamics by highlights the influence of temporal factors on buyer-seller interactions, even though the precise results are not yet available. E-commerce platforms and behavioural economics studies from the past emphasise how crucial timely negotiations are. This research could clarify the effects bargaining process delays and provide insightful guidance for practitioners and academics navigating the intricacies of online transaction. it is advised to view the entire paper in order to fully comprehend their conclusion and empirical support.

Mallik *et al.* (2023) investigate the use of neutrosophic logic for predictive analysis in recommend-er systems, contrasting it with different deep learning models. The study probably explores how well neutrosophic logic, a mathematical framework for dealing with uncertainty, performs in contrast to more conventional deep learning techniques. This work adds to the rapidly developing field recommend-er systems by examining novel approaches

to improve prediction accuracy and offering perspectives on the possible benefits and constraints of applying neutrosophic logic in this particular setting. Having access to the entire proceedings would provide a thorough comprehension of their comparative analysis.

Al-Qudah *et al.* (2023) compared the e-commerce sites of eBay and Amazon, two major players in the market. Examining different aspects like features, buyer-seller relationships, and user interface, the study sheds light on the unique tactics that each platform uses. The study's objective is to determine the platforms' functional strengths and weaknesses with a particular focus on the pharmaceutical industry. Competitive studies of the largest e-commerce companies are essential for comprehending market dynamics and platform optimization for particular industries. For a thorough understanding of their findings in the context of pharmaceutical e-commerce, it is advised to view the entire article. A cost evaluation framework for software testing fault prediction techniques is presented by Behera *et al.* (2020). The research probably presents a framework for determining how cost-effective fault prediction techniques are. The authors make a valuable contribution to the improvement of software testing procedures by exploring the complex relationship between prediction accuracy and related costs. It is likely that this framework will help practitioners make better decisions and allocate resources for software quality assurance. It is advised to view the entire proceeding for in-depth understanding.

The research by Fathalla *et al.* (2020) focuses on deep end-to-end learning for used item price prediction. It is likely that the study investigating the use of sophisticated neural network architectures for thorough price prediction. The study may aid in the optimization of predictive models for the pricing of used items by utilising end-to-end learning. The application of deep learning methods represents a step forward in tackling the intricacies linked to pricing dynamics in the used goods market. Researchers and practitioners working in the fields of pricing strategies and predictive analytics can benefit from the paper's insights. A study on improving profit through deep neural network based stock price prediction by Abrishami *et al.* (2019) is presented. The study probably explores the use of deep learning methods to stock price prediction, highlights the possible influence of these methods on profitability and financial decision making. Through their investigation of the relationship between financial tools and artificial intelligence, the writers add to the current conversation about utilising cutting edge technologies to enhance stock market predictions. Cai *et al.* (2018) explored the complex architecture of mechanisms that use principles of reinforcement learning to improve decision-making in e-commerce environments. The authors hope to enhance user interactions and possibly boost user satisfaction, engagement, and system performance by incorporating reinforcement mechanisms. The study, which is presented in the conference proceedings, looks

at how e-commerce tactics and mechanisms are changing and provides insights into online platforms and reinforcement learning interact.

The research by Kalaivani *et al.* (2021) demonstrates. When buying a phone, people look at feature like memory, display, battery life, and camera quality. But, they often can't choose right because they lack tools to check prices. That's where our machine learning model comes in, built using key phone feature data. The model helps predict a phone's price range. We've used three machine learning systems: Support Vector Machine (SVM), Random Forest Classifier (RFC), and Logistic Regression. It categorizes prices as low, medium, high or very high. We got our data from one source: the kaggle platform. We improved how well we could group by using a unique feature selection method: Chi-Squared. We started with 21 features, but only 10 were essential: RAM, pixel height, battery power, pixel width, weight, internal memory, screen width, talk time, front camera and screen height. Before tweaking, the accuracy for SVM, RFC and Logistic Regression was 95%, 83% and 76%. After, these figures rose to 97%, 87% and 81%. Our trials found SVM performed the best out of the three.

This report by Pant *et al.* (2021) explores how machine learning can estimate mobile device costs, based on their feature. The goal is to simplify the model while ensuring precise pricing. Despite mentioning the relevance of price forecasting and recognizing other research, the paper needs more focus on past studies. Specific research, the techniques employed, data and accuracy levels need more representation. This will bolster the context and showcase the uniqueness of the methods used. The section detailing data gathering and examination provide a basic overview of the collated features and statistics. Missing, however, are the specifics of the data supply, volume, likely biases and which data-refining or feature-creation methods were used. Detailing these aspects will heighten the clarity and repeatability of the paper.

Using the K-Nearest Neighbors (KNN) machine learning technique, the paper tests various K values but lacks a thorough explanation of the model choice process and fine-tuning of parameters. This added explanation would help others understand why this approach was taken.

All in all, the report puts an exciting use of machine learning in mobile device cost prediction. Yet, increasing the depth of the literary review, examining the data and justifying the methods, could significantly boost its overall quality and relevance in the field.

Chapter 3

Problem Statement

In the dynamic and competitive e-commerce landscape, accurately predicting and optimising mobile phone prices remains a complex challenge for businesses.

Traditional pricing methods often fail to capture the subtle factors influencing prices, such as ever-changing customer preferences, technological advancements, and fierce competition.

This hinders businesses from:

- Maximising their revenue and profitability
- Maintaining market competitiveness
- Offering competitive prices while remaining profitable
- Providing customers with optimal value and satisfaction

Therefore, we need a robust and adaptable approach to mobile phone price prediction that leverages the power of predictive analytics. This project aims to address this challenge by:

- Developing a data-driven predictive models using web scraping to gather rich data on mobile phone features, brand, and pricing patterns from eBay.
- Implementing and comparing various machine learning algorithms like Decision Tree Regression, Linear Regression, XGBoost, SVM and KNN to predict mobile phone prices accurately.
- Analysing the key features that significantly impact mobile prices through feature importance analysis.
- Utilising exploratory data analysis (EDA) to discover hidden patterns and customer preferences influencing pricing trends.

- Building a comprehensive analytical tool that goes beyond simple price prediction, providing insights to optimise pricing strategies, enhance customer experience, and support informed decision-making.

By successfully achieving these objectives, this project will empower businesses with a strategic advantage in the mobile phone market, enabling them to:

- predict and adapt to dynamic market changes
- Set competitive and profitable prices
- Enhance customer satisfaction and loyalty
- Make data-driven decision for strategic pricing optimization

Ultimately, this project aims to transform the mobile phone pricing landscape by demonstrating the power of predictive analytics in driving business success in the ever-evolving e-commerce environment.

Chapter 4

Design

The design and methodology used in the development of the data scraping, cleaning, and analysis project are presented in this chapter. Gathering mobile phone data from eBay, cleaning and preprocessing it, doing exploratory data analysis (EDA), and developing machine learning models to forecast mobile phone prices are the main objectives of this project. The design decisions and techniques applied in each project phase are described in the sections that follow.

4.1 The Gathering of Data

eBay was selected as the data source due to its wide range of mobile phone listings, which offer a rich dataset for analysis. Iteratively going through several pages of search results, the scraping process extracted pertinent data, including the product title, price, url for product and product details for each product. It succeeded with selecting the important features for the products to take them in help for predicting the mobile phone prices.

Data Preparation:

The code was developed to perform web scraping on eBay to extract details of mobile phone listings. It uses the requests library along with BeautifulSoup for HTML parsing. Within a loop ranging from page 1 to page 49, it sends HTTP requests with specific parameters to eBay's search results page for mobiles. It collect product titles, prices, and links by parsing the HTML content. The script filters out irrelevant items by excluding those labelled "Shop on eBay." Additionally, it attempts to extract product IDs from the links using regular expressions. After attempting to extract these details, it assembles a pandas DataFrame names

'sales' to organise the collected information, including titles, product IDs, and other details, and other details, and performs basic data cleaning steps like converting the 'price' column to strings, removing dollar signs, dropping duplicates, and exporting the cleaned data to a CSV file name 'Ebay_Mobiles.csv'.

4.2 Data Cleaning

The process of data cleaning tackles flaws in datasets to improve their quality for analysis. It is a multi-step procedure: filling in missing values; formatting data consistently; eliminating duplicate entries; fixing errors. When done thoroughly, these actions enhance the accuracy and significance of the insights drawn from the data. Here's an overview of the steps which were taken in this project for data cleaning:

1. Removing Unnecessary Columns:

- The process starts by dropping the 'Unnamed: 0' column, assumed to be unnecessary for analysis.

2. Checking and Handling Null Values:

- Checking for null values across columns.
- Dropping rows with null values in specific columns ('Model', 'Storage', 'Brand').
- Filling remaining null values in the DataFrame with 'NA'.

3. Cleaning the 'RAM' Column:

- Converting 'RAM' column values to string data type and standardizing entries by removing extraneous text ('RAM').
- Utilizing a function (clean_ram) with regular expressions to extract numeric values and convert them to a standardized format (e.g., converting GB to MB). Shown in Fig.4.1.

4. Cleaning the "Processor" Column:

- Replacing various non-standard entries (such as 'NA', 'NO', 'N/A', 'Other') with a standardized 'NA'.
- Employing a function (clean_processor_value) to categorize processor information into specific core types ('Octa Core', 'Hexa Core', etc.)

```
array([4.0960e+03, 8.1920e+03,          nan, 1.2288e+04, 1.6384e+04,
       6.1440e+03, 2.0480e+03, 3.0720e+03, 1.0240e+03, 2.5600e+02,
       4.0000e+00, 5.1200e+02, 3.2000e+01, 1.1264e+04, 1.5360e+03,
       2.2528e+04, 1.2800e+02, 2.0480e+04, 2.8800e+02, 6.1440e+01,
       6.4000e+01, 1.2000e+01, 5.1200e+03, 8.0000e+00, 7.6800e+02,
       4.8000e+01, 1.6500e+02, 1.6000e+01, 2.5000e+02, 2.0000e+00,
       9.6000e+01, 1.0000e+02, 3.0000e+00, 6.5536e+04, 7.0000e+01,
       5.0000e+01, 1.8400e+02, 5.2000e+01, 3.2768e+04, 1.0240e+04,
       6.0000e+01, 5.7600e+02, 1.9456e+03, 3.0720e+01, 2.9000e+02,
       7.1680e+03, 3.0000e+01, 4.0000e+01])
```

Fig. 4.1 Standardised RAM values

5. Cleaning the 'Model' Column:

- Replacing various non-standard entries (like symbols, 'N/A', 'ukn', 'Unknown', 'None', 'No', 'other') with 'NA'.

6. Cleaning Brand Column:

- Converted all brand names to lowercase for uniformity.
- Created a dictionary brand_mapping to map specified brand variations to standardised names.
- Replaces various 'NA', 'N/A', 'unknown', 'na', 'myphone', and 'Unbranded' values with 'NA' for consistency.
- Mapped specific brand variations using the brand_mapping dictionary.
- Converted the unique brand names to their standardized versions.

7. Cleaning Connectivity Column:

- Replaced 'NA', 'N/A', 'unknown', 'na', 'see specifications', and 'see description for details' with 'NA' for consistency.
- Created multiple columns based on different connectivity technologies such as '2G', '3G', '4G', '5G', 'Bluetooth', 'GPS', 'LTE', 'NFC', 'USB Type-C', 'Wi-Fi', 'Lightning', etc.
- Identify if each technology is present in the 'Connectivity' column and created separate boolean columns for each technology.

8. Splitting the Connectivity Column:

```
Index(['Name', 'ID', 'RAM', 'Processor', 'Model', 'Brand', 'Screen Size',
      'Connectivity', 'OS', 'Camera', 'Storage', 'Price', '2G', '3G', '4G',
      '5G', 'Bluetooth', 'DLNA', 'Dual-Band', 'GPRS', 'GPS', 'LTE', 'NFC',
      'USB Type-C', 'Wi-Fi', 'Lightning', 'Micro USB', 'WAP'],
      dtype='object')
```

Fig. 4.2 Multiple columns generated out of connectivity column

- Created multiple column based on different connectivity technologies to improve data analysis and understanding of device capabilities. Shown in Fig.4.2.

9. Cleaning Camera Column:

- Removed unwanted values like 'Varies by selection', 'none', 'Yes', 'na', 'VGA', 'Not Applicable', etc., replacing them with 'NA' for consistency.
- Handled a wide range of camera descriptions, standardized values to a certain extent (e.g., '50.0MP', '108MP', 'Dual 12MP Camera', etc.) for better analysis.
- Converted inconsistent camera description into a more standardized format for better comparison and analysis.

10. Cleaning Camera Specifications

Function clean_camera_spec:

- Removes non-numeric characters from the camera specification to extract megapixel(MP) values.
- Splits the camera specifications by + sign and extract MP values from the text.
- Joins the cleaned MP values to form a standardized camera specification.

Conversion to megapixels:

- Converts the extracted MP values to a 'Mega Pixels' columns in the dataset.
- Handles missing values and replaces them with 'NA'.

11. Cleaning Operating System(OS):

- Replaces various non-standard or unknown values in the 'OS' column with 'NA'.
- Categorizes different operating systems ('Android', 'iOS', 'Windows', etc.) using a function categorize_os.
- Converts all entries to lowercase for uniformity before categorization.

12. Cleaning Storage:

- Utilizes a function `clean_ram` to clean the 'Storage' Column. As it had similar characteristics as RAM column.
- Converts storage entries to numeric values for consistency and analysis.
- Addresses missing values and ensures uniformity in data representation.

13. Cleaning Screen Size:

- Replaces various non-standard or unknown values in the 'Screen Size' column such as 'NA', 'Unknown', 'other', etc., with 'NA'.
- Extracts numeric values denoting screen size (in inches) from the text data. User regular expressions to identify and extract these values.
- Converts extracted values to a uniform format, representing the screen size in inches.

14. Cleaning Price Column

Function `calculate_average`:

- Replaces commas in the 'price' column to handle values with commas.
- Splits price ranges indicated as 'low to high' and calculates the average price.
- Converts the 'price' column entries to float type and stores them in a new column named 'price_avg'.

15. Drop Columns

- Removes unnecessary columns such as 'connectivity', 'Camera', and 'Price' from the dataset.

16. Rename Columns

- Removes 'MegaPixels' to 'Camera', 'price_avg' to 'Price' for consistency.

17. Convert columns to Float Data Type.

- Converts 'Screen Size' and 'Camera' columns to float type for numerical consistency and ease of analysis.

18. Check Null Values

- Computes the percentage of missing values for each column in the dataset.

The data cleaning process followed systematic steps. It started by removing pointless columns and handling missing values, dropping rows with certain absent entries and replacing others with 'NA'. Specific columns like 'RAM', 'Processor', 'Model', and more than underwent particular cleaning to standardize inputs, replace odd entries, extract key info, and categorize data consistently. Shows in Fig.4.3.

	Name	RAM	Processor	Model	Brand	Screen Size	OS	Storage	2G	3G	...	GPS	LTE	NFC	4G LTE	Wi-Fi	Lightning	Micro USB	NAP	Camera	Price
0	S23 Ultra 5G Unlocked Smartphone QLED 7.3" 4GB, OnePlus 10T 5G 128GB	4096	Qualcomm Snapdragon	S23 Ultra 5G	unbranded	7.3	Android	128072	TRUE	TRUE	...	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	5.00E+02	127.99
1	Moonstone Black T-Mobile S23 Ultra Smartphone 7.3" 4+64GB Android	8192	Dota Core	OnePlus 10T 5G	oneplus	6.7	NA	128072	FALSE	FALSE	...	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	5.00E+02	269.99
3	SmartPhone NE S62 T-Mobile Unlocked AT&T T.M.	4096	Deca Core	S23 Ultra	unbranded	7.3	Android	65536	FALSE	TRUE	...	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	5.00E+01	109.47
5	Apple iPhone 12 64GB Factory Unlocked	3865.103093	NA	SMARTPHONE	S62	cat	4.70339	NA	128072	FALSE	FALSE	...	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	9.28E+06	79.99
7	Apple iPhone 12 64GB Factory Unlocked AT&T T.M.	4096	Hesa Core	Apple iPhone 12	apple	6.1	iOS	65536	FALSE	FALSE	...	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	1.20E+02	284.95
...
8503	Samsung Entro Flip Phone Paylo Virgin Mobile C.	4096	Quad Core	Entro	samsung	1.8	Android	128	TRUE	FALSE	...	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	8.00E+01	40
8504	Motorola Moto G 3rd Gen (Verizon) 64GB Black	2048	NA	Motorola Moto G 3rd Gen	motorola	6.072032	Android	8192	FALSE	FALSE	...	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	9.28E+06	33.2
8505	Flare Original Samsung SGH-R200r Flip GSM mobil.	8192	NA	SGH-R200r	samsung	3.152045	NA	6219.86169	TRUE	FALSE	...	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	9.28E+06	100
8506	REVVL 4 T-Mobile Cellphone (Black 128GB)	3865.103093	NA	Revvl 4	powerbookm edic	4.70339	NA	32768	FALSE	FALSE	...	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	9.28E+06	43
8507	Apple iPhone 13 Fully Unlocked Verizon T-Mobile	2048	NA	iPhone 13S	apple	5.422284	iOS	65536	TRUE	TRUE	...	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	1.20E+02	236
7453 rows x 26 columns																					

Fig. 4.3 Final variables

4.3 Modelling Data

In order to predict mobile phone prices, we experimented with a number of regression models, including K Nearest Neighbours, Random Forest, SVM, XGBoost, Decision Tress and Linear Regression.

Models are Explained in detail below:

1. Linear Regression

By assuming a linear relationship between the independent and dependent variables, linear regression forecasts and relationship between two variables. In order to minimise the sum of squared discrepancies between the anticipated and actual values, it searches for the best line . This technique predictions and analyses data trends and is used in a variety of fields,including finance and economics. It can also be used for logistic regression, which is appropriate for binary classification problems, and multiple linear regression with several independent variables.

A classic slop-intercept form is used to calculate the best-fit line in linear regression, and it is provided here.

$$Y_i = \beta_0 + \beta_1 X_i \quad (4.1)$$

Where X_i is the independent variable, Y_i is the dependent variable, β_0 is the constant intercept, and β_1 is the slope intercept Draper and Smith (2014).

This approach uses a straight line, $Y = \beta_0 + \beta_1 X$, to explain the linear relationship between the dependent (output) variable y and the independent (predictor) variable X .

One of the main methods used in predictive modelling is called linear regression. The target variable in this case, the price of mobile phones, is assumed to have a linear relationship with the input features. The goal of the model is to identify the line that best fits the data and minimises the deviation between the target variable's observed and predicted values. To train the Linear regression model, feature like RAM, Storage, Screen Size, and Connectivity attributes are used. The resulting model provides a baseline for predictive performance by capturing the linear associations between these features and the prices of mobile phones. Standard evaluation metrics, such as Mean Squared Error (MSE) and Root Mean Square Error (RMSE), are used to evaluate the predictive performance of linear regression. These metrics measure the square differences between the observed and predicted values in order to quantify the accuracy of the model's predictions. A better model-to-data fit is indicated by lower MSE and RMSE values. The performance of the linear regression model and other regression models is compared, which helps choose the best model for estimating the price of a mobile phone.

2. Decision Tree

Decision Tree are one of the most effective supervised learning techniques for classification and regression applications. Create a tree structure similar to an organisational chart. Each internal node represents an attribute test, and each leaf node (leaf node) contains a class name. Once a stopping criterion is met, such as the maximum depth of the tree or the minimum number of samples required to split a node, the training data is recursively split into a node, the training data is recursively split into subset based on attribute values. Decision Tree Regressor is a flexible mobile cost forecasting model. Additionally, feature importance analysis is often used to determine which features have the greatest impact on model predictions, providing important insights

into key variables that influence the cost of a cell point. Together with other regression models in the project, the results of the decision tree regressor complete the broader comparative analysis Witten and James (2013).

3. Random Forest

The main advantage of random forest algorithms seems to be their versatility and ease of use, which allow them to solve regression and classification problems efficiently. This method is well suited to processing complex datasets and reducing overfitting, making it a useful tool for various machine learning prediction tasks. One of its key features is that the Random forest algorithm can process datasets containing both continuous variables, such as regression, and categorical variables, such as classification. Improved performance for regression and classification tasks. Random Forests, a powerful ensemble learning technique, are essential in this mobile phone price prediction project. To diversify the modelling process, a series of decision trees are created and each tree is trained on a different subset of the dataset. This diversity increases the robustness of the model and reduces overfitting. Random forest models use features such as RAM, Storage, Screen Size, and connectivity attributes to make predictions based on the collective evaluation of multiple trees. Standard regression metrics such as mean square error (MSE) and root mean square error (RMSE) are used to evaluate the performance of random forest regression analysis. These metrics evaluate how accurately a model predicts the future by calculating the squared difference between observed and predicted values. The power of random forest models is that they capture complex relations in data and can therefore predict outcomes more accurately than single decision trees. The relative importance analysis. The random forest results add an important perspective to the overall comparative study of regression model Svetnik *et al.* (2003).

4. SVM Regression

Supervised machine learning algorithms such as Support Vector Machines (SVMs) are used for regression and classification problems. In order for SVM to effectively classify data into different classes, it must find a hyperplane in a high-dimensional space. The goal is to minimise classification error and maximise the margin, which is the distance between the hyperplane of each class and the nearest data point. SVM can solve linear and nonlinear classification problems using different kernel functions. This study used regression to model the relationship between mobile phone features and corresponding prices. The configurable machine learning methods SVM is known for its performance in applications ranging from regression to classification. In a re-

gression context, SVM looks for a hyperplane that best shows a linear relationship between the target variable (phone cost) and the input parameters (RAM, storage, Screen Size, etc.). This type of hyperplane is determined by the choice of kernel function. In this case, the kernel function will likely be "linear". Standard regression parameters such as mean square error (MSE) and Root mean square error (RMSE) are used to evaluate the regression performance of SVMs. These metrics evaluate how accurately the model predicts values by measuring the squared deviation between observed and expected values. SVM regression is particularly useful for complex relationships and nonlinear patterns in data. The evaluation results support general benchmarking and help you select the most suitable regression model for predicting mobile phone price for project. The importance of SVM models in pipeline modelling is further enhanced by their ability to process high-dimensional data and capture complex model Drucker *et al.* (1996).

5. XGBoost Regression

XGBoost is an efficient way to build supervised regression models. By understanding its objective function (XGBoost) and the underlying learner, we can derive the validity of this theorem. The objective function includes a regularisation term and a loss function. Represents the deviation Between actual and predicted values, or the difference between the model output and the actual values. In XGBoost, linear regression and logistic regression are most commonly used. XGBoost is a team learning method. Ensemble learning involve =s training and integrating individual models (called baseline learners) to product a single prediction. This study used Extreme Gradient Boosting (XGBoost) regression, an effective and popular ensemble learning technique, to predict the cost of mobile phones. As an iterative model, XGBoost creates set of weak learners (usually decision trees) and combines their predictions to create a final reliable and accurate model. This algorithm is known for its efficiency and good performance on a wide range of datasets and uses a regularisation approach to avoid overfitting. Two common Regression parameters, Mean square erroe (MSE) and Root mean square error (RMSE), are used to evaluate the predictive ability of an XGBoost regression model. These metrics evaluate how accurately a model predicts the future by calculating the squared difference between observed and expected values. The strength of XGBoost lies in its ability to process complex relationship and identify non-linear pattern in data. Feature importance analysis is often used to determine which feature have the greatest influence on a model's predictions. This process provides detailed information about the variables that affect the cost of a cell phone. The results of the XGBoost model evaluation the

benchmarking and help you choose the most suitable regression model Chen and Guestrin (2016).

6. **K Nearest Neighbors**

For classification and regression tasks, the K-Nearest Neighbors(KNN) algorithm is a commonly used machine learning technique. This is based on the assumption that comparable data points generally have comparable labels or values. The KNN method uses the entire training dataset as a reference throughout the training phase. The selected distance metrics is used, for example: Euclidean distance. Before making predictions, determine the distance between each training sample and the input data point. The system then determines the K nearest neighbours to the input data point based on these distances. To predict the value of an input data point in a regression, the mean or weighted average of the target value of its K neighbours is calculated. The project used K-Nearest Neighbour (KNN) regression as an instance-based non-parametric algorithm for predicting mobile phone prices. KNN's make prediction based on the similarity of instances in feature space. As part of the regression, the mean or weighted average of the target variable (mobile phone prices) is calculated based on the k nearest neighbours in the feature space.

Chapter 5

Implementation

To learn more about the distribution of important variables like RAM,Storage,Screen Size , and price, we ran descriptive statistics.A range of visual aids, such as bar charts, scatter plots, and heatmaps, were utilised to investigate correlations between variables and comprehend patterns present in the dataset.

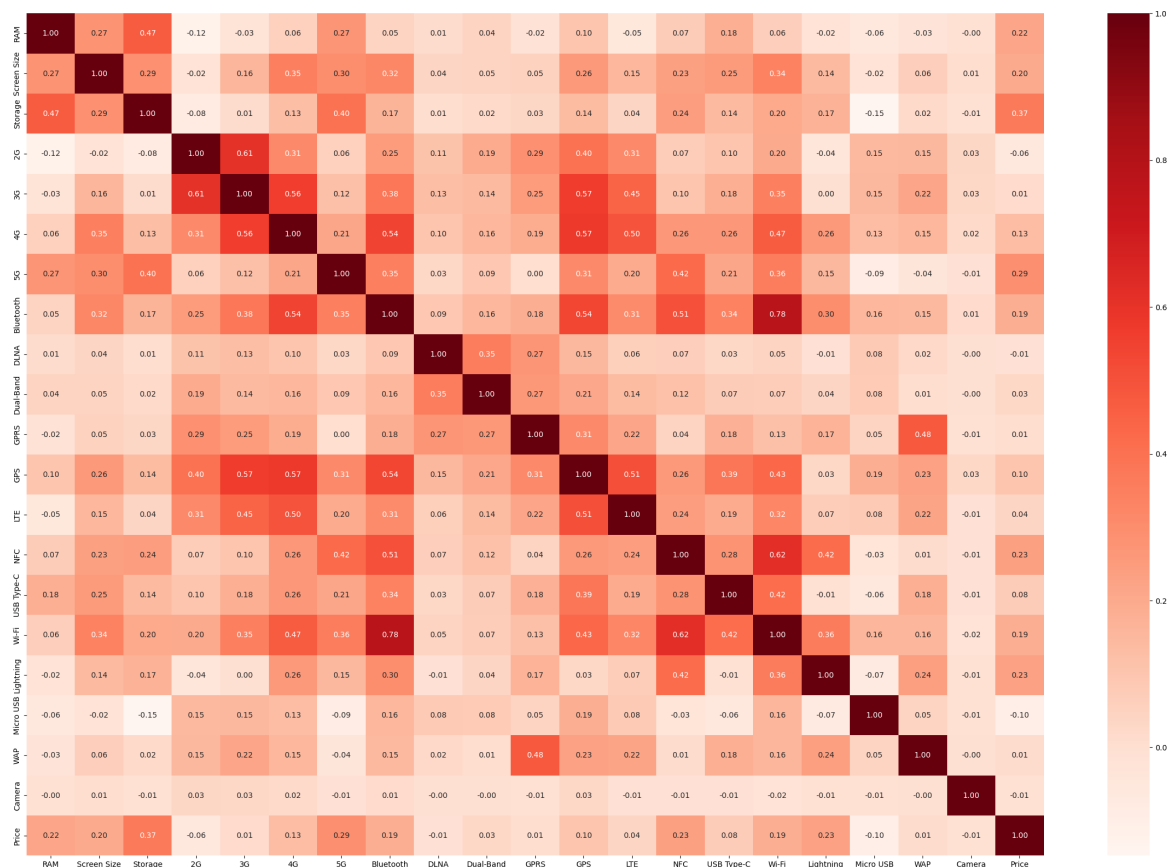


Fig. 5.1 Heat Map for sales dataset

This Heat map is used for the visual representation of correlation between the variables.

Strong positive correlations:

There are several strong positive correlations in the heatmap. For example, there is a strong positive correlation between price and RAM ,Screen Size and Storage capacity.

We also see a strong positive correlation between Bluetooth,DLNA,dual-brand and GPRS compatibility.This suggests that phones with Bluetooth compatibility are also more likely to have DLNA,Dual-brand,and GPRS support.Fig.5.1.

Negative Correlation:

There are also some negative correlations in the heatmap.For example, there appears to be a negative correlation between price and micro USB.This means that as the price of a phone increases, the possibility of micro USB decreases.

There is also a negative correlation between RAM and Micro USB.This suggests that phones with more RAM are less likely to have a Micro USB port.

Clusters and patterns:

The heatmap reveals some interesting clusters of correlation variables.For example, the variables related to price (price,RAM,Screen Size,Storage) form a tight cluster in the upper right corner.This suggests that these variables are all highly correlated with each other, and changes in one variable are likely to be accompanied by changes in the others.

Similarly, the compatibility variables (Bluetooth,DLNA,dual-brand,GPRS) form a cluster in the lower left corner, indicating a strong association between them.

The heatmap does not show any perfect correlation (1.0 or -1.0), which is to be expected in real-world data.However,the strong positive and negative correlation identified above can still provide valuable insights into the relationships between different variables.

It is important to note the correlation does not imply the actual cause.Just because two variables are correlated does not mean that one causes the other.For example, the correlation between price and RAM could be due to the fact manufacturers tend to put more RAM in higher-priced phones, but it could also be due to other factors, such as consumer demand for phones with both high RAM and high price for premium phones.

Considering the context of this industry and specific business goals can help us interpret the correlation in the heatmap and draw actionable conclusions. For example, if we are looking to increase sales of high-end phones, we might focus on promoting features like RAM, Screen Size and Storage capacity.

Overall, the heatmap is a valuable tool for exploring the relationship between different variables in the sales data. By understanding these relationships, we can gain valuable insights into the customer's preferences and make informed decisions about pricing, marketing, and product development.

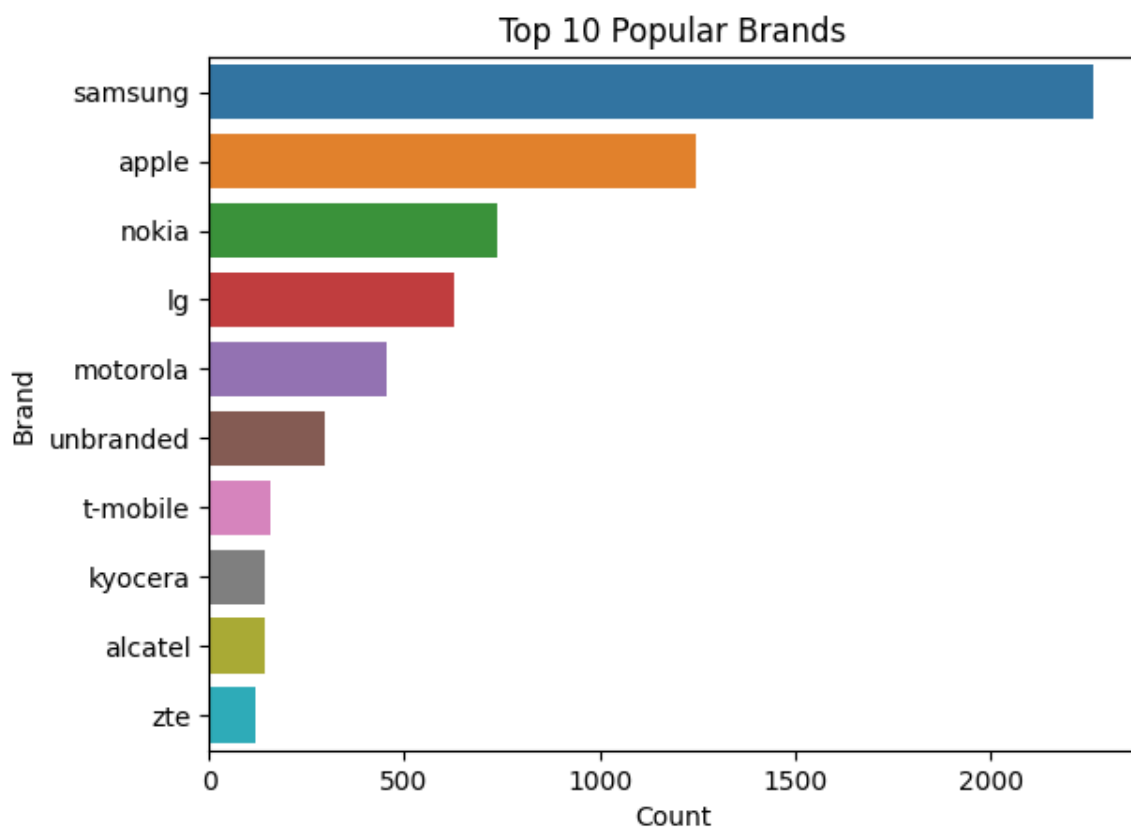


Fig. 5.2 Top 10 Brands

This graph aims to provide a visual representation of the top 10 most popular brands in the 'sales' dataset and their respective frequency of occurrence using a horizontal bar plot.

The graph shows the top 10 most popular cell phone brands.

Here are some of the key takeaways from this:

Samsung is the most popular cell phone brand, followed by Apple. This is likely due to the fact that these two brands have the largest market share as per the insights from eBay data.

Nokia is the third most popular brand, followed by LG and Motorola. These brands have been popular for many years, and they still have a loyal following.

Unbranded cell phones are also relatively popular. This is likely due to the fact that they are often more affordable than branded cell phones. Shows in Fig.5.2.

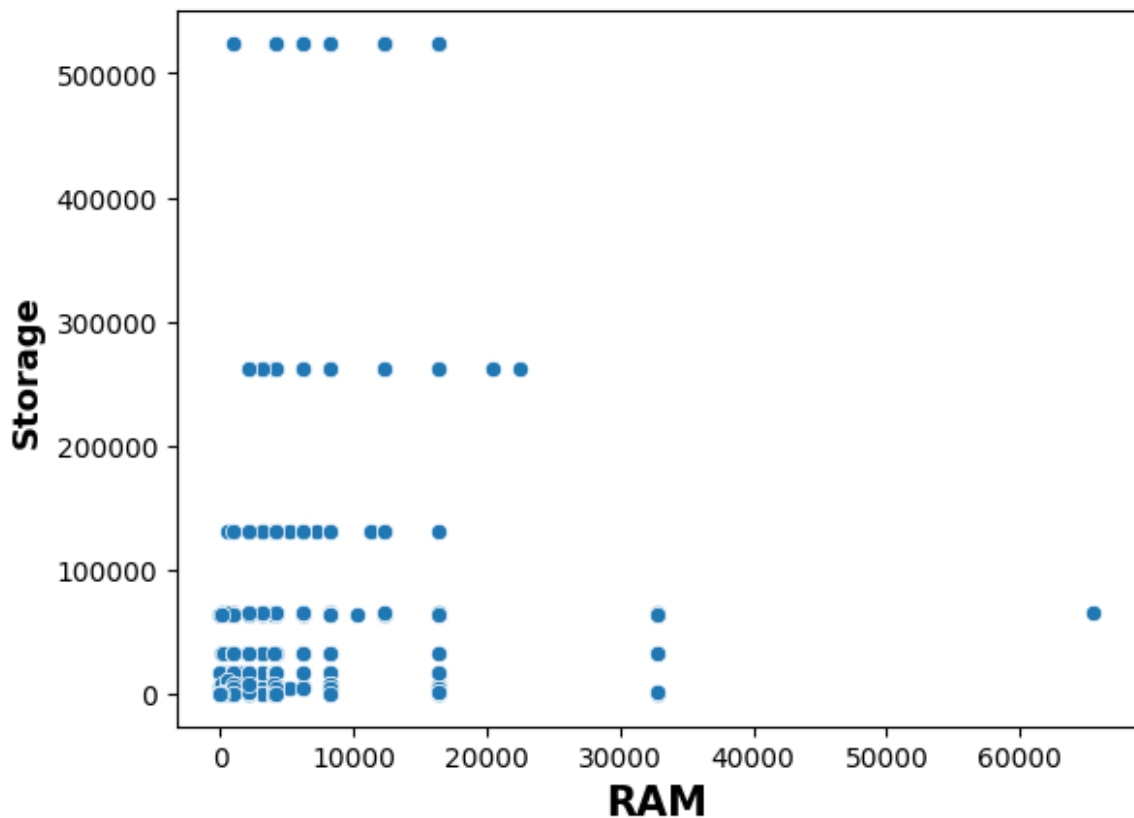


Fig. 5.3 Relationship between RAM and Storage

The scatter plot shows the relationship between storage and RAM, recorded per capita. It can be seen that there is a positive correlation between the two variables, meaning that as storage increases, RAM also tends to increase. However, there are also a few data points that fall outside of the general trends. For example, there is one data point with very high storage but relatively low RAM, this can be due to the certain models of phones which are customized with high storage and low RAM as per the user's need. Shown in Fig.5.3.

This scatter plot shows the Screen Size of the top 10 phone brands. Here are some of the key takeaways:

"Unbranded" phones shows the greatest Screen Size diversity. They range 5 inches to over 30 inches as they come from various markets. In contrast, major brands usually maintain a standard product size.

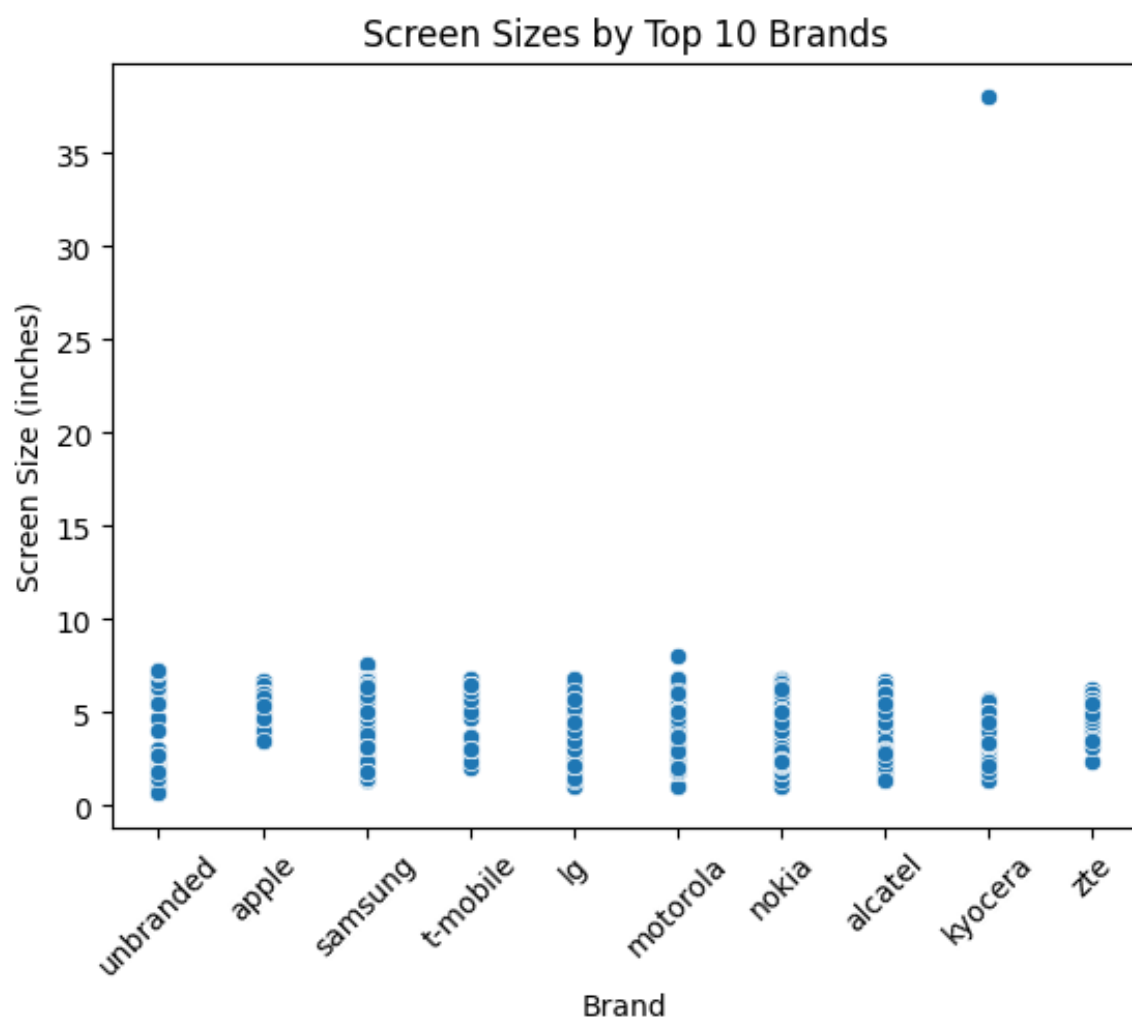


Fig. 5.4 Screen Size for top 10 Brands

Apple and Samsung provide a stable range of sizes between 5 and 7 inches. This reflects that these brands are targeting specific customers, their focus is customers wanting specific sizes. Shown in Fig. 5.4.

T-Mobile, Motorola, and Nokia exhibit greater variety in size than Apple and Samsung, yet, not as broad as unbranded phones. Likely, they attempt to serve diverse customer preferences.

Alcatel, Kyocera, and ZTE offer the smallest sizes, with their phones under 6 inches. Perhaps, their target is budget-wise buyers seeking small, low-cost phones.

Indeed, the graph uncovers a broad spectrum of available Screen Size catering to every preference. The key to phone selection lies in identifying personal requirements and Sizing needs.

Pie chart Showing the Distribution percentage of Operating system

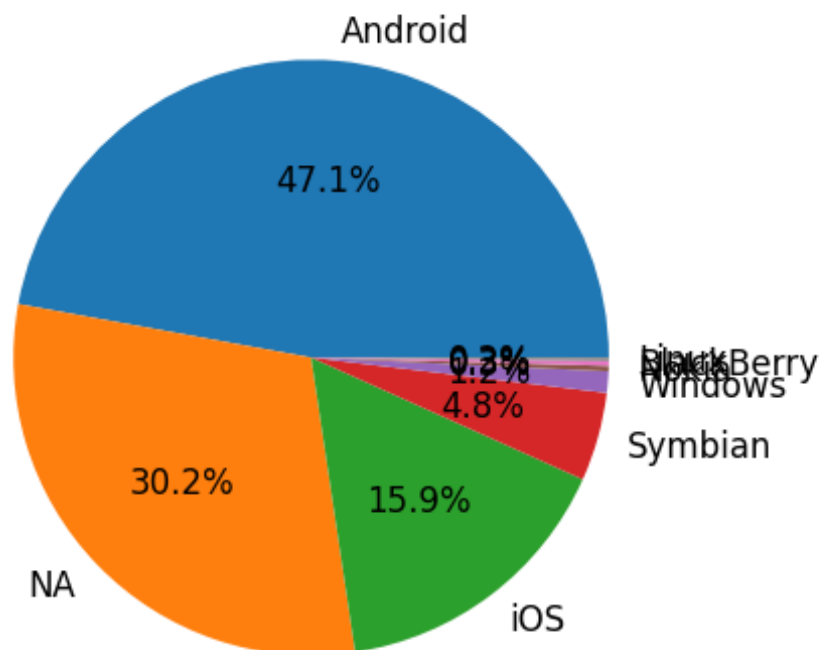


Fig. 5.5 Pie chart Showing the Distribution percentage of Operating system

Here is an analysis of the pie chart which shows the distribution of operating systems:

Android has the largest market share, with 47.1 percent. The second largest market share is with 30.2 percent which doesn't hold the information maybe because they are tagged to unbranded phones or eBay didn't provide the OS specification for these phones. Hence this can be ignored. iOS phone has the third largest market share, with 15.9 percent. Symbian has

the market share of 4.8 percent. Blackberry and other OS are smallest on the chart, they have a very small market share.

Overall, the pie chart shows that Android is the most popular operating system in the market. iOS is the second most popular operating system, followed by Symbian and Windows phone. Blackberry and Other OS have a very small market share. Shown in Fig.5.5.

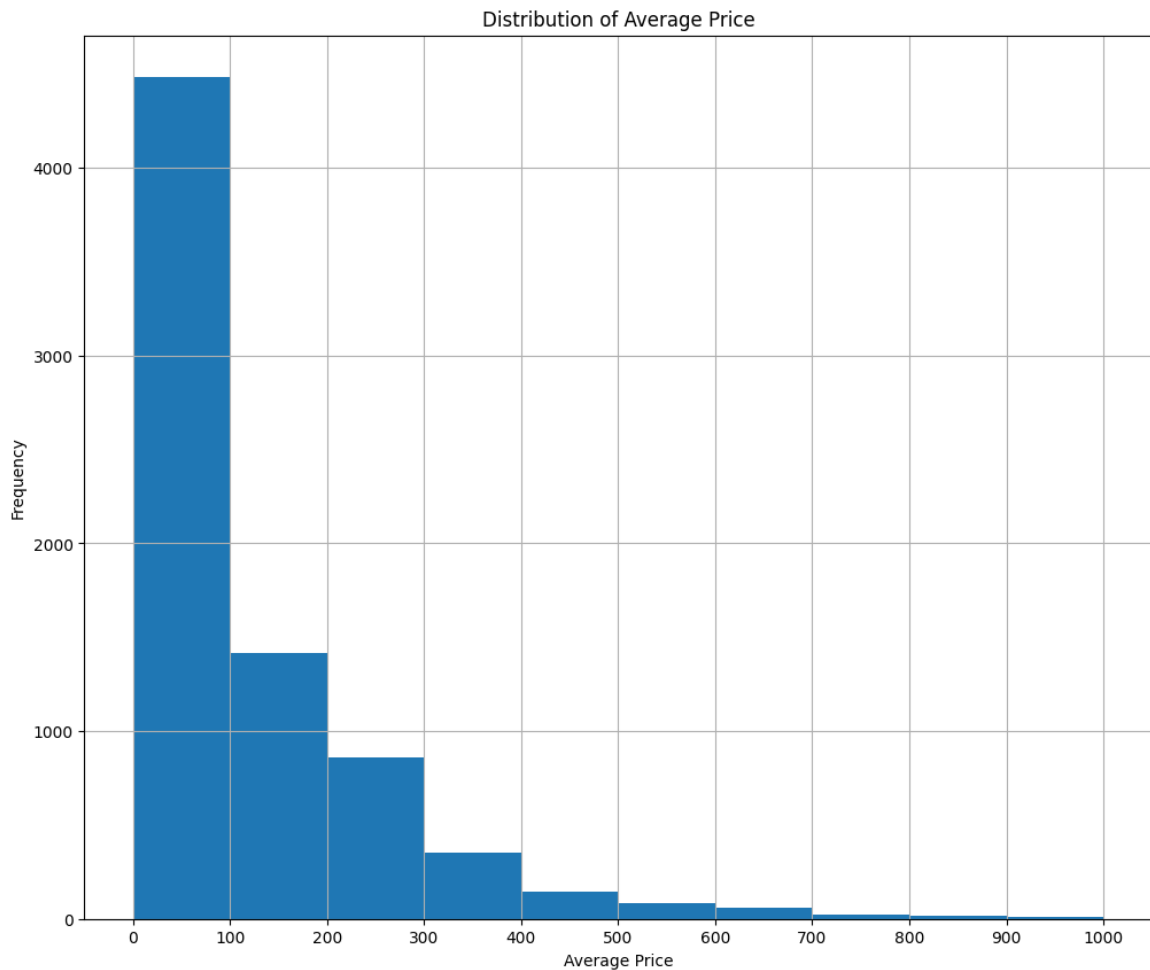


Fig. 5.6 Price Distribution

This graph displays the price distribution of the phones, and it is easy to understand that the count of low prices phones are comparatively huge in count when compared to high priced phones (premium Segments). Shown in Fig.5.6.

The graph shows the average price of smartphones by connectivity features 2G, 3G, 4G, 5G and NFC.

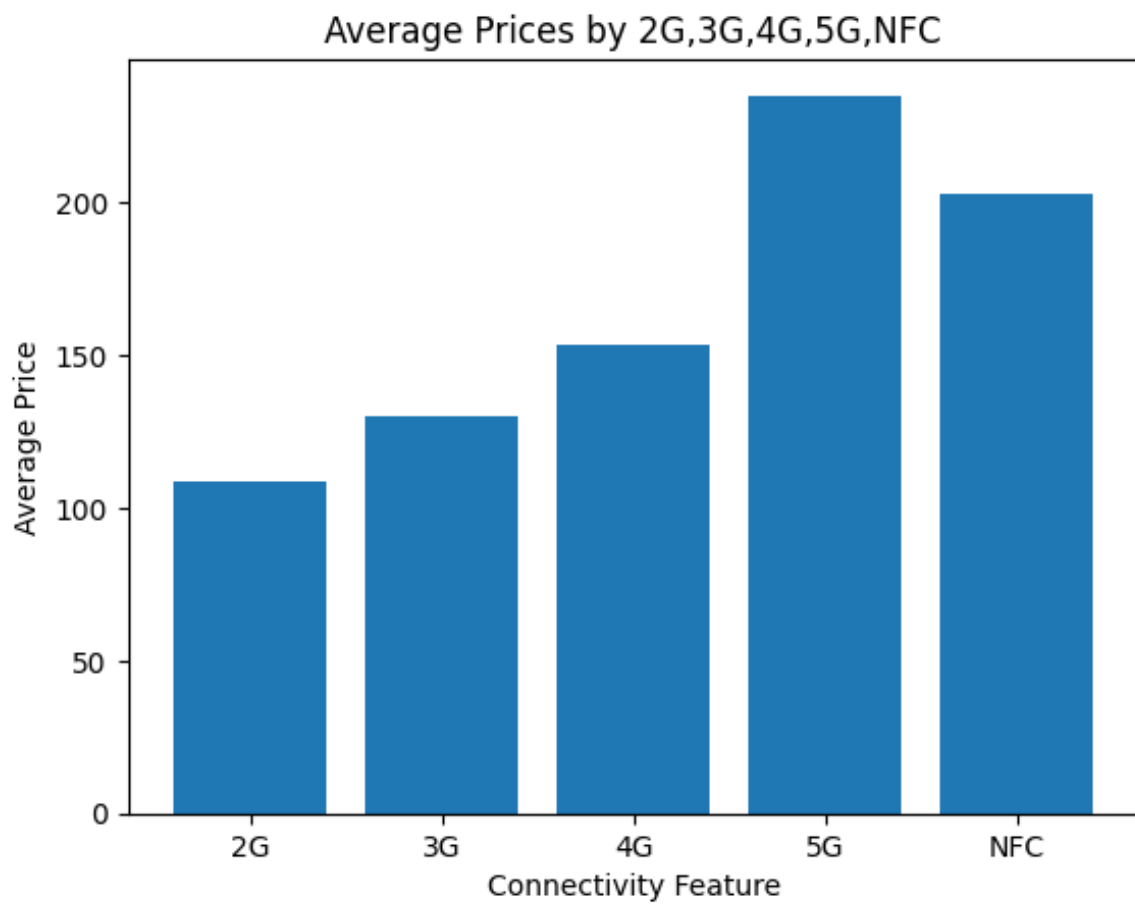


Fig. 5.7 Bar Chart for Average Price with 2G,3G,4G,5G,NFC

The average price of smartphones has been increasing steadily. This is likely due to a number of factors, including the increasing cost of components, the development of new features, and the demand for higher-end devices.

5G is the most expensive connectivity feature, followed by NFC, 4G, 3G and 2G. This is because 5G is the latest and most advanced cellular technology, and it offers a number of advantages over previous generations, such as faster speeds and lower latency.

NFC is also a relatively new technology, and it is not as widely used as 4G or 3G. However, it is becoming increasingly popular, as it allows devices to communicate with each other simply by touching them together and this usually comes bundled with 4G or 5G.

2G is the oldest and least expensive cellular technology. However, it is also the slowest and least reliable. As a result, it is no longer being used in new smartphones. Shown in Fig. 5.7.

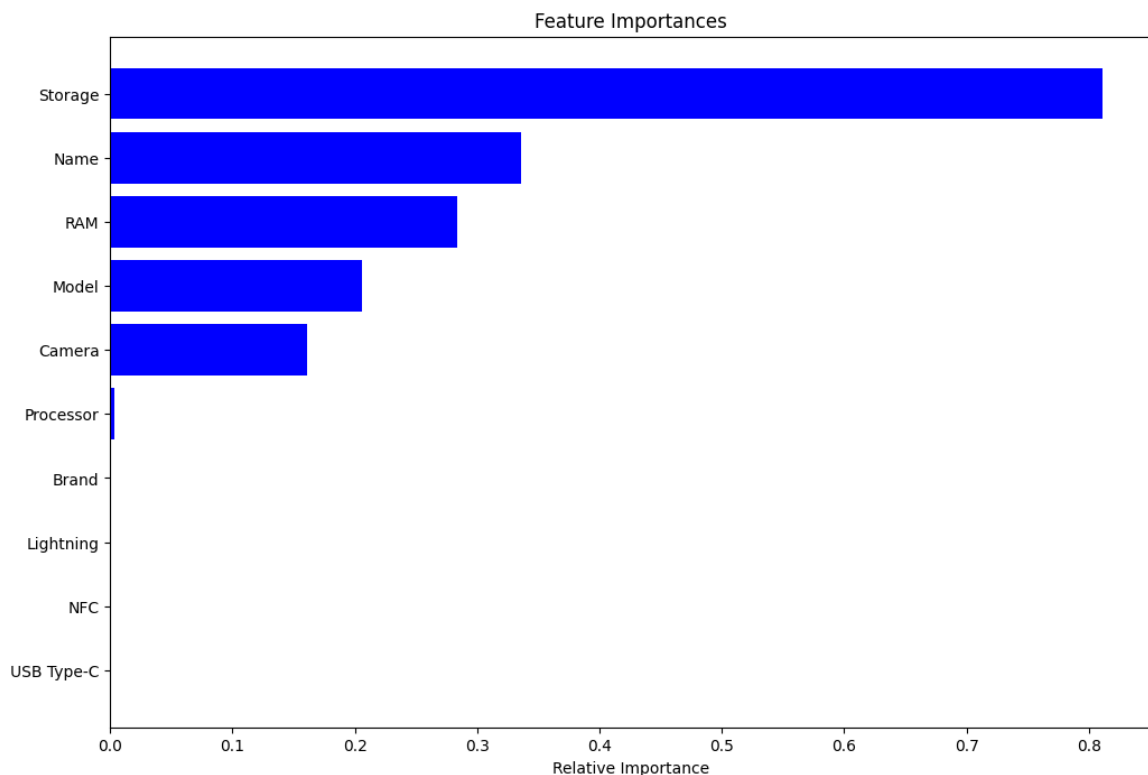


Fig. 5.8 Feature Importance

Feature Ranking:

The Graph displays the features sorted by importance, revealing the most influential features in descending order.

The top 10 feature are visualized using a horizontal bar chart which are as follow:

1. Storage
2. Name
3. Model
4. Screen Size
5. Camera
6. Brand
7. RAM
8. Processor
9. OS
10. 4G

- Relative Importance Values: The bars represent the relative importance of each feature, providing a quantitative measure of their impact.
- Feature Important: The analysis suggests that storage capacity is the most significant feature in the predictions of mobile phone prices, followed by name and model. This implies that these features play a crucial role in determining the mobile phone's price.
- Interpretations:
 1. Storage might be influential due to its impact on device functionality and user experience.
 2. Name and Model could potentially capture brand recognition, design, or specific specifications that influence user preferences.

Chapter 6

Evaluation

The predictive accuracy of the model was evaluated using performance metrics such as mean square error(MSE) and root mean square error(RMSE).

6.1 Mean Square Error:

Mean square Error(MSE) is a measure of the accuracy of a statistical model. The root mean square deviation between the predicted and observed values is calculated. A perfect model has an MSE of 0, meaning there are no errors. As a model accuracy increases, the MSE decreases, reflecting the improvement in the quality of the model predictions.

The Formula for MSE is:

$$MSE = \sum \frac{(y_i - \hat{y}_i)^2}{n} \quad (6.1)$$

6.2 Root Mean Square Error:

RMSE(Root of Mean Square Error), which represents the standard deviation of relative values, is a measure for quantifying prediction error. By considering both the residuals and the RMSE, you can evaluate the spread of these residuals and determine how much they deviate from the regression line. Essentially, RMSE provides insight into how closely your data points are clustered around a line of best fit. This indicator has application in various fields such as climatology, forecasting, and regression analysis. Additionally, it is often used to verify experimental results.

The formula for RMSE is:

$$MSE = \sqrt{\sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{n}} \quad (6.2)$$

Lower root mean square error(RMSE) and Mean square error (MSE) values indicate better model performance. With reduced MSE and RMSE values, the results indicate that the K Nearest Neighbours model performed quit well, suggesting higher accuracy in predicting mobile phone costs. Compared to linear regression, random forest , Decision Tree, SVM Regression, and XGBoost Regression also performed well, although with slightly higher error measures off. When choosing the right cell phone price prediction algorithm, you can use these metrics to provide insight into each model's predictive capabilities. Lower root mean square error (RMSE) and mean square error (MSE) values indicate better model performance. When the MSE and RMSE values decrease, the results show that the K Nearest Neighbours model perform very well, suggesting higher accuracy in predicting mobile phone costs. Compared to K Nearest Neighbours, linear regression, Random forest, decision Trees, SVM regression, and XGBoost regression also performed very well, although with slightly higher error measures. Choosing the right mobile phone price prediction algorithm can be made based on these parameters, which provide detailed information about the prediction capabilities of each model. Better forecasting performance is indicated by a decrease in the MSE/RMSE ratio. The K Nearest Neighbours model appears to have the lowest RMSE in the table provided. This suggests that it predicts the target variable (perhaps the cost of a cell phone) better than other models. K Nearest Neighbours model are very useful for dynamic pricing strategies because they can capture complex non-linear relationships and interactions between variables. They provide the flexibility needed to effectively adapt to changing consumer preferences and fluctuating market conditions. Pricing model use KNN to closely observe and analyze how different features affect prices based on similar historical data points. This approach allows companies to make informed, data driven pricing decisions based on current market trends and consumer behavior.

Predictive models, especially KNN, provide useful tools for dynamic pricing strategies. These models allows businesses to change prices in real time depending on various parameters such as demand, seasonality and competitive pricing. These models allow companies to easily adjust prices to changes in supply and demand, thereby maximising profit margins. For example, to increase sales, you can increase prices during times of high demand and offers discounts during times of low demands. Additionally, by incorporating market trends and competitive pricing into these models, businesses can conduct effective competitive analysis and make informed pricing decision to stay competitive. Predictive models enable segmented

Table 6.1 MSE and RMSE values For each Algorithm

Model	MSE	RMSE
Linear Regression	27808507342.901222	166758.8298798634
Decision Tree	16780.952509067345	129.54132583810374
Random Forest	16732.2117468851	129.35305078306078
SVM Regression	49651467725.104256	222826.09300776303
XGBoost Regression	16726.95122918556	129.3327152316287
K Nearest Neighbors	16708.26047412715	129.26043661587698

pricing strategies by identifying consumer segments and their price sensitivities, allowing companies to tailor prices to different customer groups based on preferences and willingness to pay. It is also useful for introducing. Using this comprehensive strategy, companies can effectively penetrate complex markets and make strategic pricing decisions that maintain long-term profitability and competitiveness.

Chapter 7

Conclusions

When combined with the accompanying code, the mobile price prediction model becomes an essential tool for businesses navigating the cutthroat world of e-commerce. This model creates a solid foundation for mobile phone price prediction by integrating web scraping techniques and using advanced machine learning models including models K Nearest Neighbours.

This predictive capability allows companies to develop informed pricing strategies that increase competitiveness and flexibility in response to changing market conditions. The ability to change the pricing method of this model in real time is one of its main advantages. Constantly updated with new data and adapts to variables such as demand, seasonality, and competitive pricing. This dynamic pricing strategy allows companies to maximise profit margins by adjusting prices in response to changes in supply and demand. For example, to increase sales, you can increase prices during periods of high demand and lower prices during periods of low demand. You can consider competitive pricing and market trends to make informed decision to stay competitive. Predictive models that identify customers segments and price sensitivities enable segmented pricing strategies. This personalization allows companies to adapt prices according to the preferences and willingness to pay of different customer groups. This allows businesses to increase customer satisfaction and loyalty by offering reasonable prices, targeted promotions, and customised discounts. This model goes beyond simple cost consideration to consider feature importance and brand appeal, allowing companies to make strategic decisions. Supports competitive analysis and directs joint ventures and inventory management to adapt to market trends. This Comprehensive strategy allows companies to stay focused on consumer preferences and adjust marketing and product development efforts accordingly.

Flexibility to respond to changing market conditions is one of the main motives for this research. Incorporating future upgrades provides a flexible foundation for continuous improvement, allowing businesses to quickly respond to new trends and gain a competitive

advantage in the fast-moving e-commerce space. by leveraging data-driven insights to make strategic pricing decision and improve consumer interactions, this model enable sustainable growth. Still, there is always room for improvement in the rapidly evolving fields of data science and machine learning. subsequent research could focus on improving current regression models and exploring innovative methods. There are ways to improve it, such as adjusting hyper parameters, practising regularisation techniques, and trying new methods. Future research should focus on cutting edge techniques such as deep learning models, reinforcement learning, and transfer learning. These techniques can handle complex patterns and unstructured data, improving prediction accuracy. Examining deep learning models such as neural networks and recurrent neural networks (RNNs) can reveal complex relationships in your data. In summary, businesses looking to succeed in the cutthroat world of e-commerce will greatly benefit from the mobile price prediction models. The ability to optimise profits in real time and adapt to changing market conditions positions it as a strategic tool for companies seeking long-term success. Keeping pace with developments in machine learning and data science requires continuous research and improvement.

References

- Abrishami, S., Turek, M., Choudhury, A.R. and Kumar, P. (2019) Enhancing profit by predicting stock prices using deep neural networks in: *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)* pp. 1551–1556 IEEE
- Al-Qudah, A., Al Moaiad, Y., Mohamed, R.R., Baker El-Ebiary, Y.A., Ahmad Saany, S.I., Pandey, P. *et al.* (2023) A comparative study of the e-commerce platforms of amazon and ebay. *Journal of Pharmaceutical Negative Results* **14**
- Bar-Gill, S., Brynjolfsson, E. and Hak, N. (2023) Helping small businesses become more data-driven: A field experiment on ebay Technical report National Bureau of Economic Research
- Behera, A., Das, S. and Ray, A. (2020) Cost evaluation framework for fault prediction technique in testing in: *Advances in Data Science and Management: Proceedings of ICDSM 2019* pp. 21–31 Springer
- Cai, Q., Filos-Ratsikas, A., Tang, P. and Zhang, Y. (2018) Reinforcement mechanism design for e-commerce in: *Proceedings of the 2018 World Wide Web Conference* pp. 1339–1348
- Chen, T. and Guestrin, C. (2016) Xgboost: A scalable tree boosting system in: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* pp. 785–794
- Draper, N. and Smith, H. (2014) Applied regression analysis. reprint New York: J. Wiley
- Drucker, H., Burges, C.J., Kaufman, L., Smola, A. and Vapnik, V. (1996) Support vector regression machines *Advances in neural information processing systems* **9**
- Fathalla, A., Salah, A., Li, K., Li, K. and Francesco, P. (2020) Deep end-to-end learning for price prediction of second-hand items *Knowledge and Information Systems* **62**, pp. 4541–4568
- Fong, J. and Waisman, C. (2023) The effects of delay in bargaining: Evidence from ebay Available at SSRN 4387538
- Kalaivani, K., Priyadharshini, N., Nivedhashri, S. and Nandhini, R. (2021) Predicting the price range of mobile phones using machine learning techniques in: *AIP Conference Proceedings* vol. 2387 AIP Publishing
- Mallik, S., Mohanty, S. and Mishra, B.S. (2023) Comparison of neutrosophic logic approach to various deep learning models for predictive analysis in recommender system in: *AIP Conference Proceedings* vol. 2878 AIP Publishing

- Ong, K., Haw, S.C. and Ng, K.W. (2019) Deep learning based-recommendation system: an overview on models, datasets, evaluation metrics, and future trends in: *Proceedings of the 2019 2nd international conference on computational intelligence and intelligent systems* pp. 6–11
- Pant, K., Bordoloi, D. and Pant, B. (2021) Developing a machine learning-based framework for disease prediction *Webology* **18**(5), pp. 3132–3139
- Shen, X. (2018) Predicting vulnerable files by using machine learning method
- Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P. and Feuston, B.P. (2003) Random forest: A classification and regression tool for compound classification and qsar modeling *Journal of Chemical Information and Computer Sciences* **43**(6), pp. 1947–1958
pMID: 14632445
- Szepannek, G. and Aschenbruck, R. (2019) Predicting ebay prices: Selecting and interpreting machine learning models—results of the ag dank 2018 data science competition *Archives of Data Science A (accepted)*
- Witten, D. and James, G. (2013) *An introduction to statistical learning with applications in R* springer publication
- Zhu, H. (2021) A deep learning based hybrid model for sales prediction of e-commerce with sentiment analysis in: *2021 2nd International Conference on Computing and Data Science (CDS)* pp. 493–497

Appendix A

Code Snippets and Implementation

Details

This appendix provides an insights into the technical aspects of the implementation, offering relevant code snippets that contribute to the functionality of the predictive analysis of mobile phone prices.

The following code snippets highlight key components and methods used to achieve the system's objectives.

```
URL = 'https://www.ebay.com/sch/i.html'

headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.124 Safari/537.36'}

title = []
price = []
product_link = []
shipping = []
discount = []
actual_price = []

for page_number in range(1, 50):
    params = {
        '_from': 'R40',
        '_nkw': 'mobiles',
        '_pgn': page_number,
        '_ipg': 240
    }

    page = requests.get(URL, params=params, headers=headers)

    if page.status_code == 200:
        soup = BeautifulSoup(page.text, 'html.parser')
        product_title = soup.find_all('div', class_='s-item__title')
        product_price = soup.find_all('span', class_='s-item__price')
        product_divs = soup.find_all('div', class_='s-item__info clearfix')
        for item_text, item_price, product_div in zip(product_title, product_price, product_divs):
            if item_text.text != "Shop on eBay":
                title.append(item_text.text)
                price.append(item_price.text)
                link = product_div.find('a', class_='s-item__link')['href']
                product_link.append(link)
    else:
        print(f"Error: {page.status_code}")
```

Fig. A.1 Data Scraping from eBay

```

ram = []
processor = []
model = []
screen_size = []
connectivity = []
operating_system = []
camera_resolution = []
storage_capacity = []
brand = []

for product_details in product_details_list:
    ram_value = product_details.get('RAM')
    processor_value = product_details.get('Processor')
    model_value = product_details.get('Model')
    screen_size_value = product_details.get('Screen Size')
    connectivity_value = product_details.get('Connectivity')
    os_value = product_details.get('Operating System')
    camera_value = product_details.get('Camera Resolution')
    storage_value = product_details.get('Storage Capacity')
    brand_type = product_details.get('Brand')

    ram.append(ram_value)
    processor.append(processor_value)
    model.append(model_value)
    screen_size.append(screen_size_value)
    connectivity.append(connectivity_value)
    operating_system.append(os_value)
    camera_resolution.append(camera_value)
    storage_capacity.append(storage_value)
    brand.append(brand_type)

sales = pd.DataFrame({'Name': title, 'ID': product_id, 'RAM': ram, 'Processor': processor, 'Model': model, 'Brand': brand, 'Screen Size': screen_size, 'Connectivity': connect

```

Fig. A.2 Extracting important features from the page

```

[94] from sklearn.linear_model import LinearRegression
LR= LinearRegression()
LR.fit(X_train_scale,Y_train_scale)
Y_pred = LR.predict(X_test)

```

```


mse_lr= mean_squared_error(Y_test, Y_pred)
rmse_lr = np.sqrt(mse_lr)
print("Mean Squared Error: ",mse_lr)
print("Root of mean square error:",rmse_lr)

```

```

Mean Squared Error: 27808507342.901222
Root of mean square error: 166758.8298798634

```

 Y_pred

```

array([[ 7140.59204232],
       [ 258.26203737],
       [13618.37394002],
       ...,
       [ 636.22711665],
       [27250.31909182],
       [13852.61536143]])

```

Fig. A.3 Linear Regression


```

▶ decision_tree = DecisionTreeRegressor(random_state=42)
  decision_tree.fit(X_train_scale, Y_train_scale)
  Y_pred_dt = decision_tree.predict(X_test)
  print(Y_pred_dt)

➡ [0.10451083 0.10451083 0.10451083 ... 0.10451083 0.10451083 0.10451083]

```

```

▶ mse_dt = mean_squared_error(Y_test, Y_pred_dt)
  rmse_dt=np.sqrt(mse_dt)
  print("Decision tree")
  print('Mean Squared Error:',mse_dt)
  print('Root of mean square error:',rmse_dt)

Decision tree
Mean Squared Error: 16780.952509067345
Root of mean square error: 129.54131583810374

```

Fig. A.4 Decision Tree

```

[82] random_forest = RandomForestRegressor(random_state=42)
  random_forest.fit(X_train_scale, Y_train_scale)
  Y_pred_rf = random_forest.predict(X_test)
  Y_pred_rf

array([0.3510197 , 0.34700493, 0.36683866, ..., 0.32955133, 0.36489407,
       0.37566301])

```

```

[83] mse_rf = mean_squared_error(Y_test, Y_pred_rf)
  rmse_rf=np.sqrt(mse_rf)
  print("Mean square error:",mse_rf)
  print("Root of Mean square error:",rmse_rf)

Mean square error: 16732.2117468851
Root of Mean square error: 129.35305078306078

```

Fig. A.5 Random forest

```
[4] svm = SVR(kernel='linear')
     svm.fit(X_train_scale, Y_train_scale)
     svm_pred = svm.predict(X_test)
     svm_pred

array([ 7182.80548937,   272.55307112, 13742.35212329, ...,
        640.17914123, 27509.18758532, 13942.37449105])
```

```
# Evaluate the model
svm_mse = mean_squared_error(Y_test, svm_pred)
svm_rmse = np.sqrt(svm_mse)
print(f"Mean Square Error (MSE):{svm_mse}")
print(f"Root Mean Squared Error (RMSE): {svm_rmse}")
```

```
Mean Square Error (MSE):49651467725.104256
Root Mean Squared Error (RMSE): 222826.09300776303
```

Fig. A.6 SVM Regression

```
[86] xgb = XGBRegressor()
     xgb.fit(X_train_scale, Y_train_scale)
     xgb_pred = xgb.predict(X_test)
     xgb_pred

array([0.46847245, 0.37547055, 0.4992014 , ..., 0.25294104, 0.46850795,
        0.4806715 ], dtype=float32)
```

```
xgb_mse = mean_squared_error(Y_test, xgb_pred)
xgb_rmse = np.sqrt(xgb_mse)
print(f"Mean Square Error (MSE):{xgb_mse}")
print(f"Root Mean Squared Error (RMSE): {xgb_rmse}")
```

```
Mean Square Error (MSE):16726.95122918556
Root Mean Squared Error (RMSE): 129.3327152316287
```

Fig. A.7 XGBoost

```
[88] knn = KNeighborsRegressor(n_neighbors=5)
      knn.fit(X_train_scale, Y_train_scale)
      Y_pred_knn = knn.predict(X_test)
      Y_pred_knn
```

```
array([[0.53651577],
       [0.52667882],
       [0.53651577],
       ...,
       [0.51325896],
       [0.53651577],
       [0.53651577]])
```

```
▶ # Calculate metrics
  mse_knn = mean_squared_error(Y_test, Y_pred_knn)
  rmse_knn = np.sqrt(mse_knn)
  print("Mean squared error:", mse_knn)
  print("Root mean squared error:", rmse_knn)
```

```
➞ Mean squared error: 16708.26047412715
   Root mean squared error: 129.26043661587698
```

Fig. A.8 K Nearest Neighbors

Detailed Explanation of Technical Aspects

- **Data Scraping:**The script uses web scraping to extract mobile phone information from eBay, gathering information such as titles,prices and links.
- **Data Cleaning and Preprocessing:** Several cleanup steps have been use,such as handling zero values,converting data,and standardizing values in columns such as RAM,Processor,Brand,Connectivity,Camera,OS,Storage,Screen Size etc.
- **Data Modeling and Evaluation:**Several regression models (Linear,Decision Tress,Random Forest,SVM,XGBoost,KNN)are used and their performance is evaluated using Mean Squared Error and Root of Mean squared Error.

Appendix B

Data Sources and ethical Considerations

This appendix delves into the data sources utilized throughout the project and addresses the ethical considerations and data privacy measures undertaken to ensure responsible research practices.

Data Source Used in the Project

The main data source used in the project is eBay, specifically the eBay website, from which the script extracts information about mobile phones. The script uses web scraping techniques to gather information such as title, price, contact details, and other information about mobile phones listed on eBay.

Discussion of Ethical Considerations and Data privacy

Ethical consideration and data privacy are carefully considered in the project. The researchers used web scraping techniques to collect mobile phone data mainly from eBay. Respect and adherence to the terms of use of the eBay website was important, in order to ensure that website scraping activities complied with legal and ethical standards. To prioritize usage consent, the researcher obtained explicit permission from website owners or administrators before deleting data. This option was crucial to avoid unauthorized access and potential legal consequences. A key focus of the project was the responsible handling of personal data, especially if the deleted data included personally identifiable information. The researchers used techniques such as anonymity, aggregation and data protection violations related to user privacy in order to minimize the risk of exposure. By addressing these ethical considerations and supporting data privacy principles, the researcher aimed to ensure that the businesses

complied with legal and ethical standards,respected their compliance confidentiality of the role,and promotes responsible data practices.