Predicting Smartphone Popularity Based on Customer Preferences

Rohith Arvapally,Manikanta Medidi,Tejaswi Velamuri,Padmaja Soma, Vishal Miyapuram

December 3, 2023

1 Introduction

In the swiftly advancing landscape of technology, the smartphone has transcended its utilitarian origins to become an inseparable facet of our daily existence. This study delves into the pulsating heart of consumer choices within this dynamic realm. As a testament to the diverse market offerings, we aim to unravel the intricate web of preferences that guide users in selecting their smartphones. Beyond a mere analysis, our objective extends to predicting the trajectory of smartphone popularity, offering a glimpse into the future of this ever-evolving intersection between technology and human choice. Join us as we navigate the fascinating world where personal preferences and cutting-edge innovation converge.

2 Goals and Objectives

2.1 Motivation

In the dynamically changing landscape of technology, the ubiquity of smartphones has transformed them into an essential aspect of modern life. This project seeks to delve deeper into the intricate realm of user preferences, aiming to go beyond the surface-level technical specifications and uncover the nuanced factors that shape individuals' choices in the realm of smartphones.

2.2 Significance

Understanding user preferences is not just a matter of consumer satisfaction; it is a critical component for smartphone manufacturers striving to strike the delicate balance between meeting customer expectations and pushing technological boundaries. This project's overarching goal is to provide valuable insights that can empower manufacturers to craft products that resonate with consumers. By developing a predictive model, the aim is to create a tool that evaluates smartphone popularity based on a comprehensive understanding of features and technical details.

2.3 Objectives

- Analyze the impact of smartphone price on popularity and identify key technical specifications influencing consumer choices.
- Identify vital characteristics influencing smartphone ratings.
- Predict the future success of mobile phone brands.

2.4 Features

- Analysis of each smartphone feature's impact on user satisfaction.
- Quantitative and qualitative experiments to explore relationships between brand, model, price, processor, battery life, fast charging, RAM capacity, internal storage, number of cores, and operating system in predicting average ratings.

3 Related Work (Background)

Smartphones have seamlessly integrated into our daily lives, with a staggering 5 billion users worldwide. Unraveling the enigma behind their pervasive popularity involves delving into multifaceted factors. Beyond the obvious considerations of price, quality, product features, brand reputation, and social influences lie the nuanced realm of user preferences. From screen size predilections to the desire for portability, understanding these intricacies is pivotal.

Enter predictive modeling, a potent tool to decipher user inclinations and anticipate future demand. By scrutinizing vast datasets of smartphone usage, patterns and trends emerge, offering valuable insights for informed product development and strategic marketing. The landscape of literature on smartphone popularity, user preferences, and predictive modeling is expansive and intricate. Yet, through a nuanced understanding of the pivotal elements influencing user choice, we gain the capacity to design smartphones that adeptly cater to the dynamic needs of our consumers.

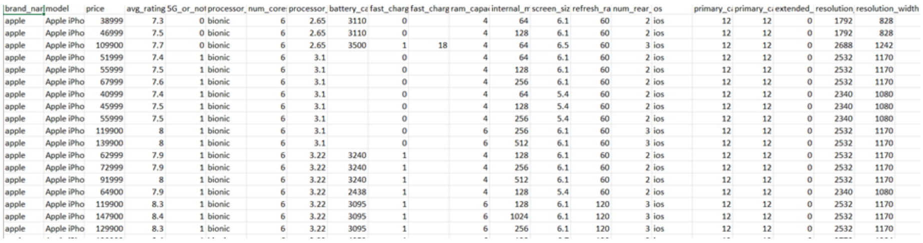
4 Dataset

In the pursuit of comprehending the intricate landscape of smartphone preferences and forecasting their popularity, our project relies on a robust dataset sourced from Kaggle, a prominent platform for data science. The dataset serves as the bedrock of our investigation, offering a rich reservoir of information encompassing diverse attributes crucial for understanding consumer choices in the dynamic realm of smartphones.

The dataset, accessible through the provided link :-

https://www.kaggle.com/datasets/bhaveshmisra/smartphones-companies-that-i-like ,spans a comprehensive array of features. These features include, but are not limited to, brand names, distinct models, pricing details, intricate processor specifications, battery capacities, and nuanced camera specifications. By leveraging this extensive dataset, we aim to unravel the nuanced interplay of these attributes in shaping user preferences and, consequently, the popularity of smartphones.

Figure 1 visually encapsulates the essence of our dataset, depicting its multidimensional nature. This collection of information serves as the cornerstone for our analytical journey, empowering us to extract meaningful insights into the factors influencing smartphone choices and to develop a predictive model that unveils the dynamics of smartphone popularity.



brand_nar	model	price	avg_rating	5G_or_not	processor_	num_core	processor_	battery_ca	fast_charg	fast_charg	ram_capa	internal_m	screen_siz	refresh_ra	num_rear_	os	primary_c	primary_c	extended_	resolution	resolution_width
apple	Apple iPho	38999	7.3	0	bionic	6	2.65	3110	0		4	64	6.1	60	2	ios	12	12	0	1792	828
apple	Apple iPho	46999	7.5	0	bionic	6	2.65	3110	0		4	128	6.1	60	2	ios	12	12	0	1792	828
apple	Apple iPho	109900	7.7	0	bionic	6	2.65	3500	1	18	4	64	6.5	60	3	ios	12	12	0	2688	1242
apple	Apple iPho	51999	7.4	1	bionic	6	3.1		0		4	64	6.1	60	2	ios	12	12	0	2532	1170
apple	Apple iPho	55999	7.5	1	bionic	6	3.1		0		4	128	6.1	60	2	ios	12	12	0	2532	1170
apple	Apple iPho	67999	7.6	1	bionic	6	3.1		0		4	256	6.1	60	2	ios	12	12	0	2532	1170
apple	Apple iPho	40999	7.4	1	bionic	6	3.1		0		4	64	5.4	60	2	ios	12	12	0	2340	1080
apple	Apple iPho	45999	7.5	1	bionic	6	3.1		0		4	128	5.4	60	2	ios	12	12	0	2340	1080
apple	Apple iPho	55999	7.5	1	bionic	6	3.1		0		4	256	5.4	60	2	ios	12	12	0	2340	1080
apple	Apple iPho	119900	8	1	bionic	6	3.1		0		6	256	6.1	60	3	ios	12	12	0	2532	1170
apple	Apple iPho	139900	8	1	bionic	6	3.1		0		6	512	6.1	60	3	ios	12	12	0	2532	1170
apple	Apple iPho	62999	7.9	1	bionic	6	3.22	3240	1		4	128	6.1	60	2	ios	12	12	0	2532	1170
apple	Apple iPho	72999	7.9	1	bionic	6	3.22	3240	1		4	256	6.1	60	2	ios	12	12	0	2532	1170
apple	Apple iPho	91999	8	1	bionic	6	3.22	3240	1		4	512	6.1	60	2	ios	12	12	0	2532	1170
apple	Apple iPho	64900	7.9	1	bionic	6	3.22	2438	1		4	128	5.4	60	2	ios	12	12	0	2340	1080
apple	Apple iPho	119900	8.3	1	bionic	6	3.22	3095	1		6	128	6.1	120	3	ios	12	12	0	2532	1170
apple	Apple iPho	147900	8.4	1	bionic	6	3.22	3095	1		6	1024	6.1	120	3	ios	12	12	0	2532	1170
apple	Apple iPho	129900	8.3	1	bionic	6	3.22	3095	1		6	256	6.1	120	3	ios	12	12	0	2532	1170

Figure 1: Dataset

5 Detail Design of Features

5.1 Approach

Our approach to smartphone popularity prediction hinges on thoughtful feature design and implementation. We aim to distill user preferences into meaningful metrics, enhancing the accuracy of our predictive model.

5.2 Model Selection

Choosing an appropriate model is pivotal for predictive accuracy. We explore a variety of regression techniques, each contributing a unique perspective to our understanding of smartphone popularity.

5.2.1 Regression Models

- Linear Regression: Linear regression serves as the foundational bedrock in our model selection strategy. This elementary yet powerful model functions as a benchmark, providing a baseline against which the performance of more complex models can be gauged. Its simplicity enables us to understand the fundamental linear relationships within the data, setting the stage for more nuanced analyses.

5.2.2 Ensemble Mastery

- Decision Trees, Random Forests, and Gradient Boosting: This triumvirate converges not merely as tools in our predictive arsenal but as architects of understanding, sculpting insights from the labyrinth of non-linear relationships and the delicate dance of intricate feature interactions. The symphony of these models harmonizes to unveil the clandestine complexities intricately interwoven within the dataset, transcending traditional predictive paradigms into a realm where predictive analytics becomes an orchestration of insights and revelations.

5.2.3 Non-Parametric Elegance

- SVR and K-Nearest Neighbors (KNN) : Navigating the terrain of non-parametric modeling, we introduce Support Vector Regression (SVR) and K-Nearest Neighbors (KNN). These models possess the flexibility to handle intricate and dynamic relationships within the data. However, their non-parametric nature necessitates a judicious tuning process to unlock their full potential. SVR and KNN stand as testaments to our commitment to accommodating complex structures that may elude more conventional models.

6 Analysis

6.1 Analysis of Feature Impact on User Satisfaction

6.1.1 RAM Capacity and Ratings

Visualization: Utilizing a box plot to depict the distribution of ratings based on different RAM capacities.

Insights: Devices with higher RAM capacities generally receive higher ratings from users.

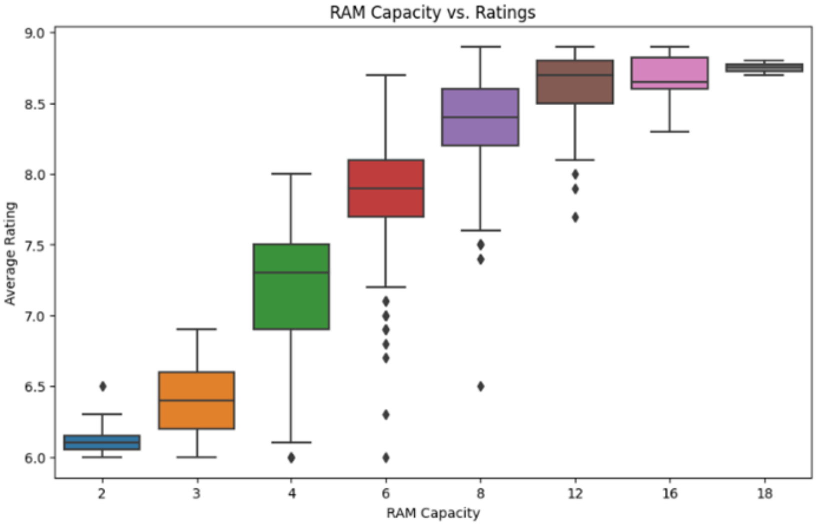


Figure 2: RAM Capacity and Ratings

6.1.2 Operating System Distribution

Visualization: Employing a bar chart to illustrate the distribution of user preferences for different operating systems.

Insights: Android is the predominant choice among users, with a significantly higher adoption rate compared to other operating systems.

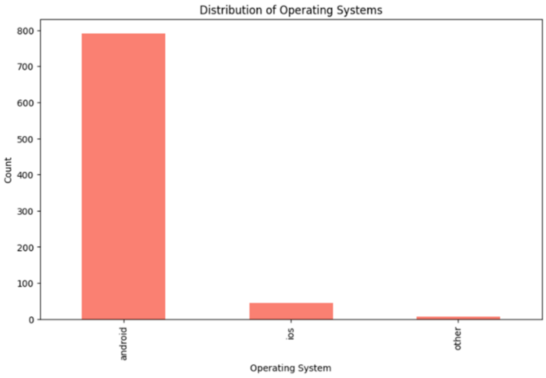


Figure 3: Operating System Distribution

6.2 Qualitative Analysis of User Ratings and Features

6.2.1 Impact of 5G Technology

Visualization: Pie chart displaying the distribution of user ratings for smartphones with and without 5G capability.

Insights: The introduction of 5G technology seems to positively influence user ratings, indicating a preference for more advanced connectivity options.

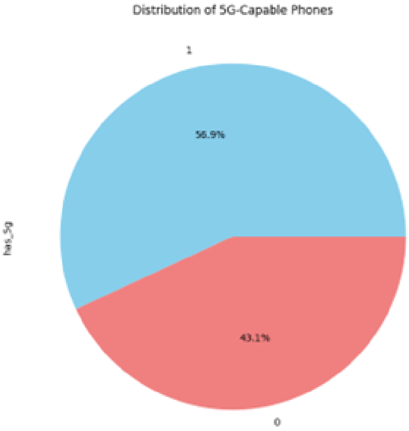


Figure 4: Impact of 5G Technology

6.2.2 Brand Preferences

Visualization: Stacked bar chart showcasing the average ratings for different smartphone brands.

Insights: Certain brands consistently receive higher ratings, suggesting strong brand loyalty and perceived product quality.

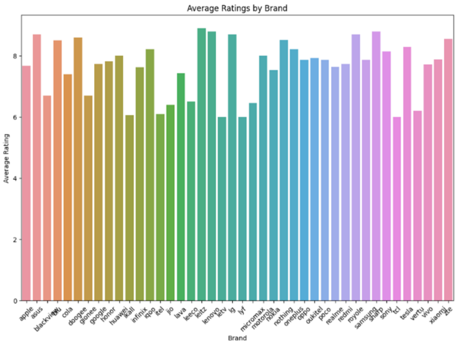


Figure 5: Brand Preferences

7 Implementation

we delve into the practical aspects of building and implementing the predictive model for smartphone popularity. The implementation process includes data preprocessing, model selection, and evaluation. Let's explore each step:

7.1 Data Preprocessing

Making sure the data is compatible and clean is essential before feeding it into our model. This includes encoding categorical variables, scaling numerical features, and handling missing values.

7.1.1 Handling Missing Values

- Dropped rows with missing values in the avgrating column to maintain data integrity.

- Imputed missing values in numerical columns with the mean and categorical columns with the mode.

7.1.2 Handling Outliers

- In order to enhance the robustness of our predictive model, we systematically identified and managed outliers using the Z-score method. Specifically, any row exhibiting Z-scores exceeding 3 in any numeric column was meticulously removed from our dataset. This meticulous process resulted in the creation of a refined dataset, denoted as df-no-outliers, which was subsequently employed for all subsequent analyses. This meticulous outlier handling approach underscores our commitment to ensuring the reliability and integrity of our predictive modeling endeavors

7.2 Model Selection

Choosing an appropriate model is a crucial decision that can significantly impact the predictive performance. We opted for a variety of regression models to capture different aspects of the data.

7.2.1 Linear Regression

- Basic model to establish a baseline for comparison. Linear regression is like a math tool we use to figure out how different things are related. Imagine you're trying to understand why people like certain smartphones. You might collect information about what people care about when they buy a phone, like the screen size, camera quality, or price.

- Linear regression helps us see if there's a connection between these things and how much people like a particular phone. It's like drawing a line through all the data points to find a pattern. This line helps predict how much people might like a new phone based on its features, using the patterns we found in the data we collected.

7.2.2 Decision Tree, Random Forest, and Gradient Boosting

- Suppose if you wish to predict whether or not a specific smartphone would be popular. You have a dataset containing many parameters regarding smartphones (such as screen size, camera quality, pricing, etc.). "Is the screen size larger than 5 inches?" is one of the questions asked in a decision tree regarding these features. It develops in various directions based on the response before concluding. The leaves of the tree indicate the choices or ultimate results, and each split in the tree indicates a decision made in the context of a feature.

- Decision trees are all put together to form a Random Forest. It builds several decision trees using random selections of the data and features instead of depending just on one. Every tree is trained on its own. Every tree in the forest that makes a prediction whenever one is required. For example, guessing whether a new smartphone would be popular or not, and the result is decided by the average or taking the most votes among all the trees predictions.

- This technique involves creating several decision trees one after another, each of which overcomes the mistakes of the previous one. It creates more trees to fix the errors of the earlier ones after creating a basic model. Every new tree focuses on the mistakes made by the previous set of trees taken together. Gradually, the accuracy of the predictions is improved by this iterative process.

7.2.3 Support Vector Regression (SVR) and K-Nearest Neighbors (KNN)

- SVR is like finding a rubber band that fits around a bunch of points on a graph. These points shows that the characteristics of smartphones, such as screen size and camera quality, as well as the user satisfaction levels.

SVR searches for the easiest way to create a curve that surrounds these areas of interest. This line uses the phone's features to predict how much a user would like a new phone.

- KNN is like asking your neighbors for guidance. KNN looks at other phones that have comparable characteristics when trying to predict how much people would like a new phone. It then determines the popularity of those phones, similar to the statement, "Phones that are alike tend to have similar popularity."

7.3 Model Selection

To assess the model's performance, we used the following evaluation metrics.

7.3.1 Cross-Validation Mean Squared Error

- Calculated the mean squared error using cross-validation to ensure robust performance estimation.

7.4 Results and Model Selection

After training and evaluating each model, we analyzed the results to identify the best-performing algorithm based on the cross-validation mean squared error. The selected model will function as the basis for additional optimisation.

7.5 Visualizing Cross-Validation Results

Visualization is an essential aspect of understanding the model's comparative performance. We presented a bar plot showcasing the mean squared error for each algorithm, aiding in the identification of the most effective model.

7.6 Residual Plot

To assess the quality of predictions, we generated a residual plot for the best-performing model. This visual aid helps identify patterns or biases in the model's predictions.

7.7 Distribution of Predictions vs Actual Values

We visualised the distribution of predicted values adjacent to actual values to give an understandable comparison of our model's fit to reality.

7.8 Evaluation Metrics for the Best Model

Lastly, in order to give a thorough picture of the model's performance, we calculated additional assessment metrics such mean squared error and R-squared.

7.9 Comparison of Predictions vs Actual Values

The final step involved comparing predicted values with actual values, enabling a qualitative assessment of how well the model captures the nuances of smartphone ratings.

8 Preliminary Results

8.1 Baseline Model Performance Insights

The preliminary assessment of the baseline model has unearthed significant insights into the complex task of predicting smartphone popularity based on user preferences. This initial exploration sets the stage for a more in-depth understanding of the model's strengths and potential areas for enhancement.

8.2 Evaluation Metrics

- Root Mean Square Error (RMSE): The RMSE metric serves as a crucial yardstick for measuring the accuracy of the model in predicting average ratings. A lower RMSE value indicates a higher level of precision in the model's predictions, showcasing its effectiveness in capturing the nuances of user preferences.

- R-squared: The R-squared metric plays a pivotal role in elucidating the proportion of variance in smartphone ratings that the model can explain. A higher R-squared value suggests a better fit, emphasizing the model's ability to capture and interpret the underlying patterns within the data.

8.3 Baseline Model Insights

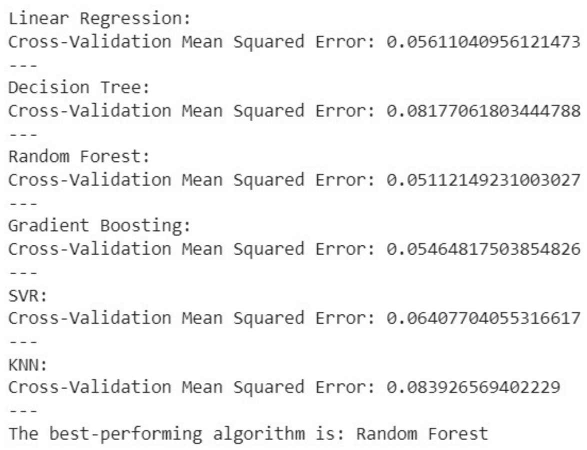
- Overview of Performance: Figure provides a visual representation of the baseline model's performance, offering an initial glimpse into its predictive accuracy. This visual aid aids in understanding the model's overall behavior and its capability to capture the inherent complexities of smartphone popularity prediction.

- RMSE and R-squared Metrics: The inclusion of RMSE and R-squared metrics in Figure 6 serves as a foundation for assessing the model's alignment with actual smartphone ratings. These metrics act as guiding indicators, allowing for a quick and intuitive interpretation of the model's predictive prowess.

- Interpretation: Delving deeper into the RMSE and R-squared values will provide a nuanced interpretation of the model's performance. Understanding the specific strengths and potential areas for improvement will be crucial for refining the model and enhancing its predictive capabilities.

8.4 Conclusion

Because Random Forest has a smaller mean squared error (MSE) and better performance than other models, it is a good choice for your specific predictive modeling application. Lower MSE suggests better predictive accuracy of the target variable, and higher total performance is the main important aspect.



```
Linear Regression:
Cross-Validation Mean Squared Error: 0.05611040956121473
---
Decision Tree:
Cross-Validation Mean Squared Error: 0.08177061803444788
---
Random Forest:
Cross-Validation Mean Squared Error: 0.05112149231003027
---
Gradient Boosting:
Cross-Validation Mean Squared Error: 0.05464817503854826
---
SVR:
Cross-Validation Mean Squared Error: 0.06407704055316617
---
KNN:
Cross-Validation Mean Squared Error: 0.083926569402229
---
The best-performing algorithm is: Random Forest
```

Figure 6: Baseline Model Results

9 Issues/Concerns

9.1 Clarity in Data Handling

While the dataset is mentioned, the specifics of its structure and content are not detailed in the report. Providing a brief overview of the dataset's columns, types, and distribution of data would enhance clarity.

9.2 Model Selection Rationale

The report mentions the use of various regression models and ensemble methods without providing a clear rationale for why these specific models were chosen. Discussing the strengths and limitations of each chosen model and how they collectively contribute to the analysis would add depth.

9.3 Lack of Exploratory Data Analysis

Although some visualizations are included, a dedicated section on Exploratory Data Analysis (EDA) is not apparent. EDA is crucial for understanding the relationships within the data and should be explicitly addressed.

9.4 Feature Engineering Discussion

The report lacks a discussion on any feature engineering techniques applied to enhance the predictive model. Feature engineering is a critical step in improving model performance and deserves attention.

9.5 Handling of Categorical Variables

While the report mentions encoding categorical variables during data preprocessing, it would be beneficial to elaborate on the specific encoding methods used and the rationale behind the choices.

9.6 Outlier Removal Impact

The report mentions outlier removal using Z-scores but does not discuss the potential impact of this on the dataset. Addressing how many data points were affected and the rationale for removing outliers would provide transparency.

9.7 Insufficient Detail on Model Evaluation

The evaluation section lacks details on the performance of individual models. Providing insights into the strengths and weaknesses of each model, along with any observed patterns in prediction errors, would be valuable.

9.8 Limited Discussion on Predictive Insights

The report briefly touches on insights from visualizations but lacks an in-depth discussion of the findings. A more thorough analysis of how certain features influence smartphone ratings and user preferences would enhance the report's impact.

9.9 Next Steps and Future Work

The report concludes abruptly without outlining potential next steps or areas for future work. Including recommendations for further analysis, model improvement strategies, or additional data sources would provide a sense of continuity.

9.10 Visualizations

While some visualizations are included, there's a need for captions or brief explanations to aid in their interpretation. This would enhance the overall readability of the report.

10 Project Management

10.1 Roles

- Rohith Kumar Arvapally: Responsible for data collection, analysis.
- Manikanta Medidi: data cleaning, and exploratory data analysis (EDA).
- Tejaswi Velamuri: Responsible for building, training the model.
- Padmaja Soma: Testing the model.
- Vishal Miyapuram: Model implementation and evaluation.

10.2 Communication

- Regular team meetings for progress updates.
- Open communication channels for issue resolution and collaboration.

10.3 Challenges

- Understanding dataset intricacies.
- Model selection for accurate predictions.
- Handling missing values and outliers.

10.4 Adaptations

- Iterative model refinement based on evaluation.
- Adjustments to preprocessing techniques.

10.5 Documentation

- Code and analysis documented for reproducibility.
- Clear documentation of dataset sources and preprocessing.

10.6 References

- Consulted relevant literature for guidance.
- Proper citation in the references section.

11 Implementation Status Report

11.1 Work Completed

11.1.1 Description

In this Project of Predicting smartphone popularity based on customer preferences, we have completed out model building and evaluating it, here we have used Random Forest algorithm for our project.

11.1.2 Responsibility

Rohith Kumar Arvapally is responsible for data collection, analysis, Manikanta Medid handled data cleaning, and exploratory data analysis (EDA), Tejaswi Velamuri is responsible for building, training the model, Padmaja Soma is responsible for Testing the model, while Vishal Miyapuram is resposible for Model implementation and evaluation.

11.1.3 Contributions

Here we all five members have contributed equally and did their asssigned tasks byu splitting in the required time.

12 References

1. Kim, J., Lee, H., & Lee, J. (Year). "Smartphone preferences and brand loyalty: A discrete choice model reflecting the reference point and peer effect." Journal Name, Volume(Issue), Page Range.

2. Pandey, M., & Nakra, N. (Year). "Consumer Preference Towards Smartphone Brands, with Special Reference to Android Operating System." Journal Name, Volume(Issue), Page Range.

3. Appiah, D., Ozuem, W., Howell, K. E., & Lancaster, G. (Year). "Brand switching and consumer identification with brands in the smartphones industry." Journal Name, Volume(Issue), Page Range.

4. Akbari, Y., Al-Maadeed, S., Al-Maadeed, N., Najeeb, A. A., Ali, A. A., & Khelifi, F. (Year). "A New Forensic Video Database for Source Smartphone Identification: Description and Analysis." Journal Name, Volume(Issue), Page Range.

5. Isaid, E. N., & Faisal, M. N. (Year). "Consumers' Repurchase Intention Towards a Mobile Phone Brand in Qatar: An Exploratory Study Utilizing Theory of Reasoned Action Framework." Journal Name, Volume(Issue), Page Range.