Understanding Mobile Phone Features Importance in India

Adam Hall, Mohamed Hussien, Virginia Sahagun, Nazmus Sakib Sumon, Shadman Chowdhury

Georgia Institute of Technology

Introduction

The behavior of consumers toward smartphone features is increasingly becoming a key focus of marketing research. Even more so in India, as the Indian smartphone market is one of the most competitive smartphone markets in the world. It is projected that it will grow 10 % in 2023 to reach 175 million units [1], with the country having almost 600 million smartphone users [2]. It is currently the second largest smartphone market in the world. With this high number of smartphone users, we also have a market, where there is not a specific brand ruling the market (Unlike USA where Apple has a large market share). In the last 5 years new smartphone brands such as Xiaomi, Vivo, RealMe have risen to take a substantial portion of the smartphone market. This is only possible due to the different types of consumer markets that exists within India. Different users have different needs; thus, the marketing mix is a key component to get popularity in this market. Along with this upward trend of smartphone users, the increasing popularity of e-commerce sites such as Flipkart has created an opportunity for new brands of smartphones to penetrate this market with the right product and marketing mix. This creates a need for quantitative research that explores what are the key factors that are believed to pique consumer interest the most.

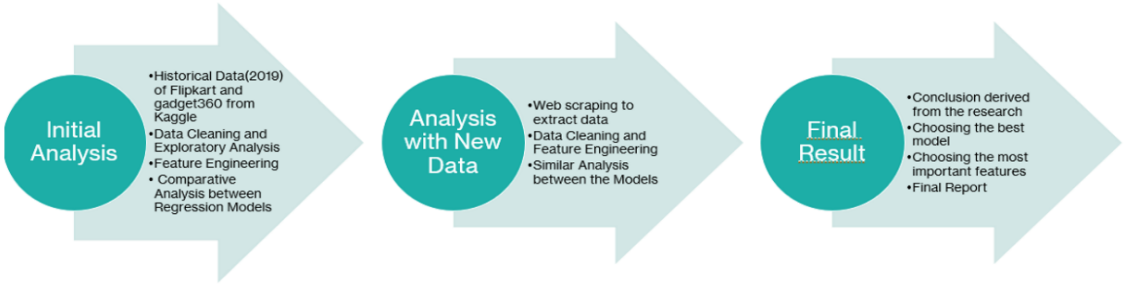
Problem Statement and Value proposition

As stated, there is a necessity to have quantitative research that can help new brands penetrate the user market, by helping them understand which factors influence a buyer's buying decision the most. The problem we found was that there is little research that has been done recently on this particular topic. With the rise of new smartphone brands such as Xiaomi, Realme etc. it is necessary to revisit the market to understand what features impact the buyer's choice the most. We plan on exploring this by using two different datasets, one from the e-commerce site Flipkart and another from Gadget360.

The dataset from Flipkart will give us some basic smartphone information such as their price, camera specifications ratings etc. We augment each record by joining this dataset with Gadget360 dataset, and having more specification information for a particular mobile device such as processor, RAM, ROM, battery etc. Based on the joined dataset, we plan on using regression-based Models such as Linear Regression, XGBoost Regression, Random Forest Regression to find out how well our model is determining the rating of the mobile phone based on a specific configuration of a mobile phone. We also want to see what are the features that impact the ratings of the smartphone the most.

Additionally, we want to see if the trends we see in our datasets match historical trends. For that we have also gathered the 2019 version of the Flipkart and Gadget360 Dataset where we will perform the same analysis to see what are the main features that impact the rating the most. This way we can see if there is a historical trend in the most popular smartphone features. A high-level diagram of our work process is given below:

Planned Approach



Initial Hypothesis

Based on a preliminary analysis from consulting reports and existing data analysis on cell phone usage and market trends, our initial hypothesis is that consumers are in-tune and highly aware of cell phone features and factors that contribute to a more convenient and efficient user experience and are likely to result in higher popularity among cell phones. Specifically, features such as larger screens, longer battery life, faster processing speeds, and improved camera quality are likely to be attractive to users who value functionality and ease-of-use. Additionally, factors such as affordability, brand reputation, and marketing efforts may also play a significant role in determining a cell phone's popularity. However, it is important to note that the relative importance of these factors may vary across different demographic groups- but we are focused exclusively on the Indian market.

Methodology

The aim is to predict customer phone ratings in the Indian market based on multiple phone features. To achieve this, a combination of Linear Regression, PCA + Linear Regression, Ridge Regression, Lasso Regression, Random Forest, and XGBoost will be used. We will begin with Linear Regression to establish a baseline model. This method will give us an idea of the most significant phone features that contribute to customer ratings. However, in our exploratory data analysis (EDA) we found that our data exhibits multicollinearity, therefore we will also perform Linear Regression with the outputs of Principal Component Analysis (PCA). Additionally, PCA will help to reduce the dimensionality of our data.

We will also use Ridge Regression. This method reduces the overall fitness of the model without eliminating factors, which will reduce the likelihood of overfitting without the risk of removing potentially important phone features. Like Ridge Regression, Lasso Regression also places a constraint on the sum of coefficients, however it removes the least important factors allowing for feature selection. Lasso regression also reduces the likelihood of overfitting.

We will also use Random Forest and XGBoost, two decision tree methods. In our literature review, we found that XGBoost created a good model for a similar analysis using app features to predict app ratings [4]. Additionally, XGBoost models tend to do well when there are complex feature interactions in the data.

Model Evaluation. We will evaluate the performance of each model using accuracy, mean squared error (MSE), and adjusted R-squared. By comparing the performance of each model, we will be able to identify the best method for predicting phone popularity in the Indian market.

Data sources and cleaning:

The dataset of this project is composed of 2 complementary datasets. A dataset from Flipkart.com which includes mobile phones ratings and some phone features e.g., price and RAM, and a complementary dataset from Gadgets360.com which provides more phones features that are not available in the Flipkart dataset.

Given the rapidly evolving nature of the mobile phone market, data sourcing passed through two phases that resulted in 2 datasets that represent different snapshots of history:

a) Existing (2021) datasets: datasets available on Kaggle for both Flipkart and Gadgets360 website. Despite the datasets being readily available, they both date back to 2021.
b) New (2023) datasets: Web scrapped datasets of the websites of Flipkart and Gadgets360 website using Python to get 2023 view of the market.

Web scrapping challenges:

The web scrapping of Gadgets360 had some challenges. The landing page for Gadgets360 mobile data contains only a few specifications for each mobile. If someone needs to see all available features, one needs to click the link for that mobile and see the specifications. To solve this issue, we used a for loop to iterate over these web pages and to get the necessary specifications for each mobile. The webpages contain sponsored mobile advertisements, and the different ways of presenting the specification on these advertisements were producing bad outputs. So, we had to update the code to avoid these advertisements. We had to avoid specifications of other products available on the webpage, like laptops, cameras, tablets, etc., which were being shown on some webpages.

One problem was deciding whether we should keep all specifications for all mobiles or decide to remove some “rare” specifications which were not found in most mobiles. We knew that we would do cleaning and merging later, still including all these specifications of all mobiles, would only unnecessarily increase column numbers. So, those “rare” specifications were not included. It can be seen as a limitation of our web scrapping.

The web scrapping of Flipkart data was like that of Gadget360. We had to do string manipulation to clean the web data. Some other modifications were made, like removing the currency sign from the price data and removing ‘mp’ from camera quality information.

Final Dataset

Although a lot of data cleaning was conducted during web scrapping of the new datasets, additional data cleaning steps were required as described in the table below.

Step	Gadgets360	Flipkart
Initial dataset size	8,695 rows x 50 columns	1,998 rows x 17 columns
Remove useless columns	8,695 rows x 6 columns	1,998 rows x 13 columns
After missing values dropped	7,800 rows	1,998 rows
After duplicates dropped	4,268 rows	1,040 rows
Merging	599 rows x 17 columns	

We kept most of the columns of Flipkart and planned to add additional columns from Gadget360. A lot of the columns had the same value for different mobiles. Like, all mobiles have these exact specifications, which will not add value to our analysis. And some columns were giving data already present in Flipkart data. That's why we had to remove a lot of columns, and in the end, it had only six columns.

Initial merging attempts of these two datasets only gave a handful of rows of matching columns. After some research, we found that leading and trailing spaces on the values of joining columns were causing the mismatch between these two datasets. We also converted the values into uppercase to make the values similar. It did work, and we got the matching columns.

Finally, string columns (e.g., operating system) are unified and converted to factors and Release date was converted to days since release. The final dataset is composed of 3 factor-based columns and 12 numerical columns in addition to the dependent variable (rating). The Summary of the dataset is described in the appendix.

Exploratory Data Analysis

For the numerical columns, a correlation matrix was constructed to give an insight into the correlation between the rating of the phone and against predictor variables as well as evaluating the correlation between predictor variables.

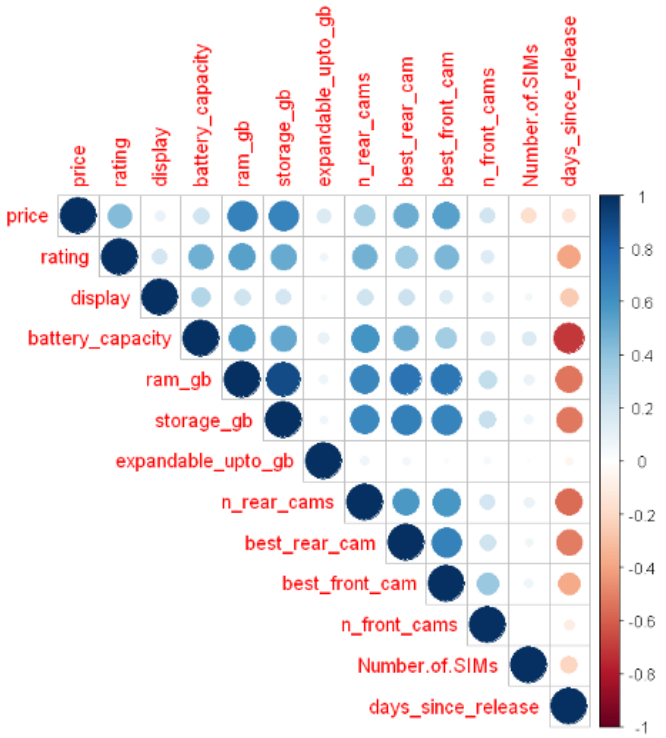


Figure 1. *Correlation matrix for phone features from the merged dataset*

From the correlation matrix, it can be deduced that:

1. Rating is directly correlated with battery capacity, RAM, storage capacity and cameras and inversely correlated with days since release, i.e., older phones have lower rating. (Aligned with initial hypothesis)
2. Some predictor variables (such as cameras and price, RAM and storage capacity, etc.) are well correlated which might require regularized regression to reduce the impact of collinearity. This is further supported by the Variance Inflation Factor (VIF) shown in the appendix.

For the categorical (factor-based) variables, most of the dataset is dominated by Android based touch screen phones (582 touch screen phones and 585 Android phones out of 600 phones). There are 42 different brands in the dataset with the leading brands being Vivo, Oppo and Asus.

Analysis of the dependent variable, rating, showed the presence of 1 outlier with rating 2.2. Investigating the outlier, it was found that this phone had only 1 review despite being released over 5 and half years ago and thus it was safe to omit the data point and assume that it was not representative. The distribution of ratings and its QQ-plot is shown in the figures below.

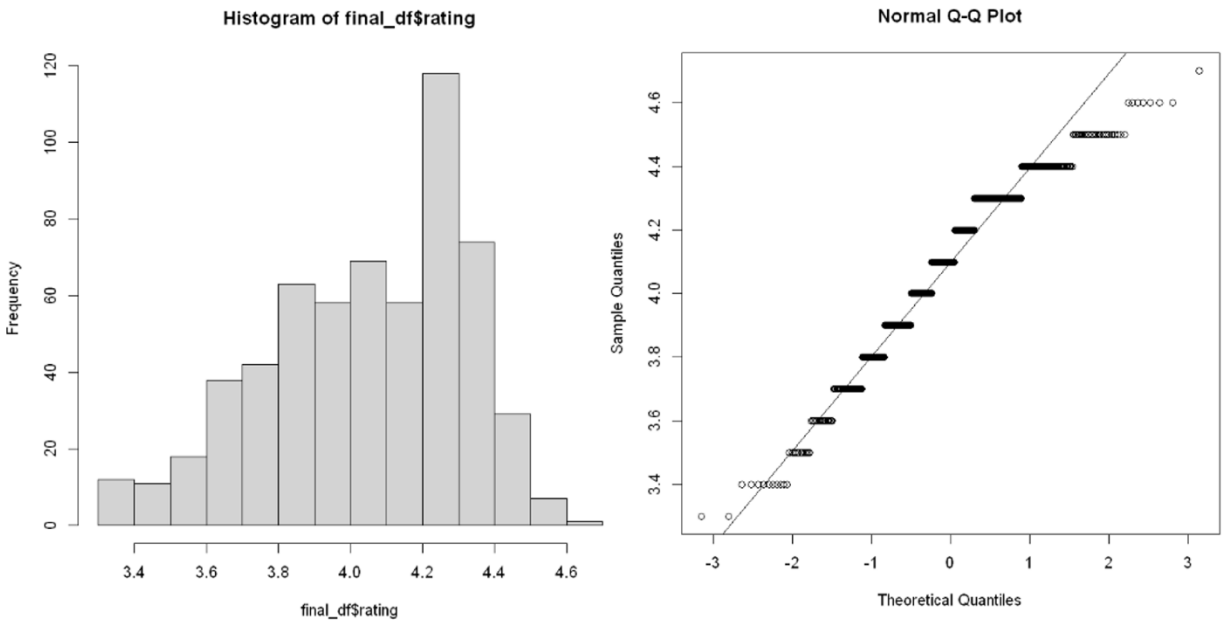


Figure 2 *(Left). Histogram of phone ratings.* ***Figure 3*** *(Right). Q-Q plot comparing data to a Normal Distribution.*

From the figure above, rating is not normally distributed and is rightly skewed. This is mainly due to the common rating of 4.2. Simple variable transformation will not be able to correct the skewness of the data.

The Novelties in our approach and initial findings

The uniqueness of our approach stems from the following three key factors-

1) Our Focus in Indian Market: Unlike other parts of the world, the Indian market is really varied when it comes to Smartphone Brands. The Indian market is also a representation of the wider south Asian smartphone markets. A lot of the brand that gains popularity here, eventually

finds popularity in other neighboring countries such as Bangladesh, Pakistan etc. Also since India is the second largest smartphone market in the world, a model created on the information extracted from this market, can be a representation of the market in the other neighboring countries as well.

2) Historical Trends- Another novelty in our approach is that we are not only focusing on the snapshot of the current market for our analysis. Rather we are also doing the same analysis on an older version of the datasets to see if there is a correlation between our findings from each of the two different timeliness
3) Augmentation of the data records of the ecommerce site- The data records found in the ecommerce site Flipkart, albeit especially useful, can be enhanced by adding more data for each of the phones that exist in the dataset. For that we took the base of the phones that were found in the original Flipkart dataset, and joined that with the dataset of the gadget360, where there is a comprehensive specification information for different smartphone devices. By joining these two datasets, we felt like we have a lot of variables to work with, that will help in our analysis of the features through the different machine learning models.

The result of our initial experiment showed that there is a high correlation between the different features of the smartphones. Which is expected, as the normal convention states that, increase of price will result in better specification. Based on the joined datasets from the 2019 versions, we saw that the main features that impact the rating the most are Brand, Price and Storage capacity. This aligns with our initial hypothesis. Due to the nature of the correlation between the features, we cannot say this result is exhaustive. So, we plan on reducing the correlation between the features, perform better feature extraction through PCA and then train our model to see if the initial result of our experiment still exists in that modified dataset as well. Based on our results from this version of the dataset, we want to implement the same process in the other joined dataset to see if the results align with our findings in that dataset as well.

Literature Reviews

According to research by Pekka in phone popularity on features from 2004-2013, most changes in phone popularity are related to the technical evolution of phones: their display, communication, and camera capabilities. However, around 2007 there was a shift as software became a larger aspect of consumers' buying decision, showing a fast-moving dynamic in consumer trends. More research was conducted by Odaymat in 2019, showing memory capacity and price as the leading factors impacting phone popularity. Odaymat sampled twenty of the most popular phones, a ran a regression model to find which factors were significant in predicting popularity. The two studies demonstrate that the value in extracting how features impact popularity, as consumer desires and hence sales can quickly shift.

References

[1] B. Standard, "Business Standard," [Online]. Available: https://www.business-standard.com/article/technology/india-smartphone-market-to-grow-10-to-reach-175-million-units-in-2023-122122300733_1.html.

[2] T. E. Times, "The Economic Times|News," 26 October 2021. [Online]. Available: https://economictimes.indiatimes.com/news/india/indias-growing-data-usage-smartphone-adoption-to-boost-digital-india-initiatives-top-bureaucrat/articleshow/87275402.cms?from=mdr.

[3] Pekka Kekolahti, Kalevi Kilkki, Heikki Hämmäinen, Antti Riikonen, Features as predictors of phone popularity: An analysis of trends and structural breaks, Telematics and Informatics,Volume 33, Issue 4, 2016, Pages 973-989, ISSN 0736-5853, https://doi.org/10.1016/j.tele.2016.03.001.

[4] https://medium.com/analytics-vidhya/is-it-possible-to-predict-rating-of-google-play-store-apps-based-on-the-given-information-da9a44a3ac1e

[5] Odaymat, Rommy. (2019). Factors affecting mobile phone selection. https://www.researchgate.net/publication/333461879_Factors_affecting_mobile_phone_selection

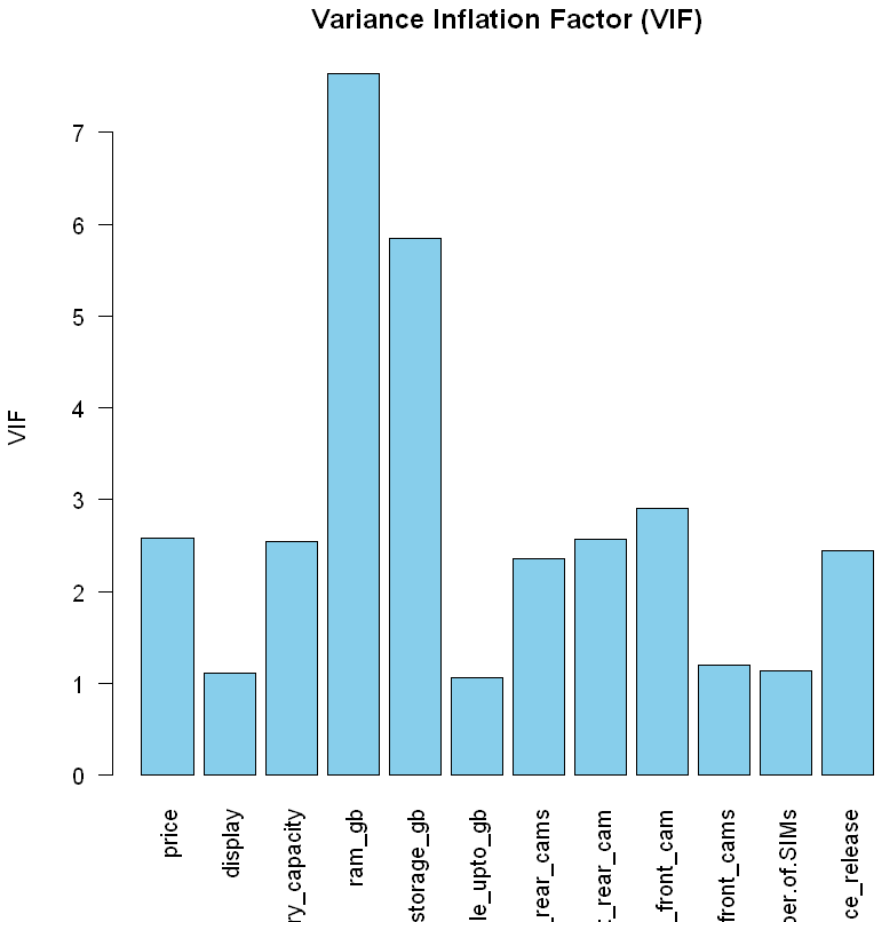
Appendix

Summary of Final Dataset

```
      brand         price            rating         display
 VIVO     : 86   Min.   :  1489   Min.   :2.300   Min.   :  4.50
 ASUS     : 49   1st Qu.:  8499   1st Qu.:3.900   1st Qu.: 13.21
 OPPO     : 49   Median : 12499   Median :4.100   Median : 14.83
 SAMSUNG  : 38   Mean   : 16813   Mean   :4.092   Mean   : 14.90
 NOKIA    : 36   3rd Qu.: 19990   3rd Qu.:4.300   3rd Qu.: 16.51
 PANASONIC: 35   Max.   :128999   Max.   :4.700   Max.   :167.64
 (Other)  :306
 battery_capacity     ram_gb          storage_gb      expandable_upto_gb
 Min.   :1020     Min.   : 0.004   Min.   :  0.004   Min.   :    0.0
 1st Qu.:3000     1st Qu.: 2.000   1st Qu.: 16.000   1st Qu.:   32.0
 Median :3500     Median : 3.000   Median : 32.000   Median :  128.0
 Mean   :3692     Mean   : 3.782   Mean   : 61.469   Mean   :  374.6
 3rd Qu.:5000     3rd Qu.: 4.000   3rd Qu.: 64.000   3rd Qu.:  256.0
 Max.   :7000     Max.   :18.000   Max.   :512.000   Max.   :20000.0

  n_rear_cams    best_rear_cam    best_front_cam   n_front_cams
 Min.   :1.000   Min.   :  0.30   Min.   : 0.00   Min.   :0.0000
 1st Qu.:1.000   1st Qu.: 12.00   1st Qu.: 5.00   1st Qu.:1.0000
 Median :1.000   Median : 13.00   Median : 8.00   Median :1.0000
 Mean   :1.761   Mean   : 20.72   Mean   :10.35   Mean   :0.9983
 3rd Qu.:2.000   3rd Qu.: 16.00   3rd Qu.:13.00   3rd Qu.:1.0000
 Max.   :4.000   Max.   :200.00   Max.   :60.00   Max.   :3.0000

 Operating.system Touchscreen Number.of.SIMs  days_since_release
 Android:585      No : 17     Min.   :1.000   Min.   :  13.0
 OTHER  : 14      Yes:582     1st Qu.:2.000   1st Qu.: 843.5
                              Median :2.000   Median :1734.0
                              Mean   :1.957   Mean   :1660.8
                              3rd Qu.:2.000   3rd Qu.:2372.0
                              Max.   :3.000   Max.   :3256.0
```


Variance Inflation Factor (VIF)
VIF
0
1
2
3
4
5
6
7
price
display
ry_capacity
ram_gb
storage_gb
le_upto_gb
_rear_cams
:_rear_cam
_front_cam
front_cams
ber.of.SIMs
ce_release