Understanding the Importance of Mobile Phone Features in India

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

Georgia Institute of Technology

Introduction

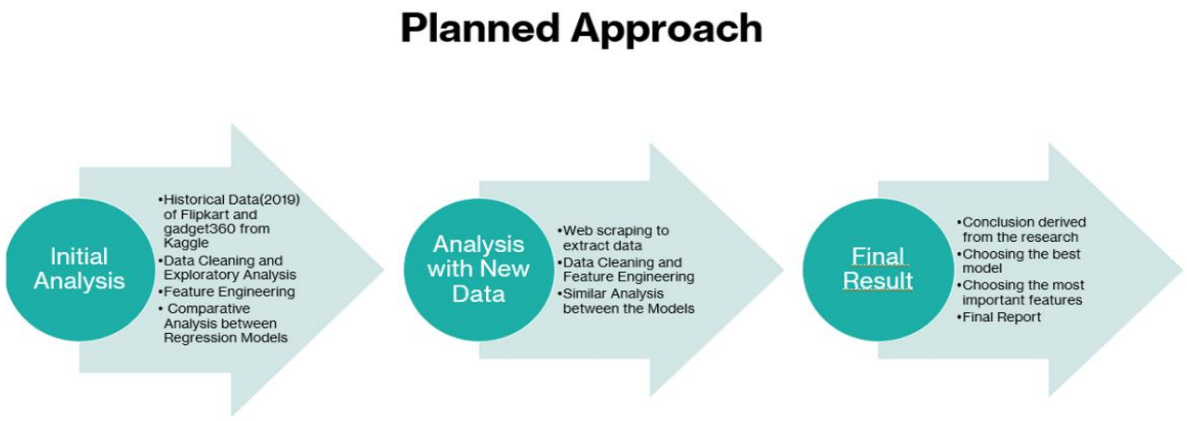
Consumers are becoming more aware of the features and value provided by their cell phone. With the available options growing every year, companies can benefit by understanding the customers better and delivering on the key features that are important to them. Even more so in India, as the Indian smartphone 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]. Along with this upward trend of smartphone users, the increasing popularity and availability of e-commerce sites such as Flipkart has created an opportunity for customers to easily compare smartphones. [3] New or existing brands may look for quantitative research that explores the key factors that are driving consumers' interest.

Problem Statement and Value Proposition

As stated, it is necessary to have quantitative research that can help new and existing brands grow their market share by helping them understand which factors influence a buyer's decision the most. The problem we found was that there is little research that has been done recently on this topic. With heightened competition, it is necessary to revisit the market frequently to understand what features impact the buyer's choice the most. We plan to explore 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 that has more specification information for mobile devices such as processor, RAM, ROM, and battery. Based on the joined dataset, we plan to use regression models to see which features and aspects of the phones are useful in predicting their ratings and which of those models is best in explaining their ratings.

Additionally, we want to see if the key features we see in our datasets are consistent over time. For that we have also gathered the 2019 version of the Flipkart and Gadget360 Dataset where we will perform the same analysis to find which features impact ratings the most. This way we can see if there are trends in the most popular smartphone features. A high-level diagram of our work process is given below:



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.[4] 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 and ran a regression model to find which factors were significant in predicting popularity. The two studies demonstrate that phone features can drive ratings and popularity, but those features have changed over time.[6]

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 features such as larger screens, longer battery life, faster processing speeds, and improved camera quality are likely to be attractive to consumers. These consumers are in-tune and highly aware of cell phone features that contribute to a more convenient and efficient user experience and are likely to result in higher popularity among cell phones. 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.

Data Sources

The dataset of this project is composed of two complementary datasets from Flipkart.com and Gadgets360.com. The dataset from Flipkart.com includes mobile phone ratings and 15 different, but basic, features such as price and RAM. Gadgets360.com is a complementary dataset that provides an additional 40 phone features that were not available in the Flipkart dataset. Given the rapidly evolving nature of the mobile phone market, data was gathered in two points in time. There were differences in column numbers in the two versions of the data, but the key features we wanted to focus on existed in both. The two different versions are-

a) Existing (2019) datasets: Available on Kaggle for both Flipkart and Gadgets360 website.

b) New (2023) datasets: Web scrapped datasets of the websites Flipkart and Gadgets360 using Python to get 2023 view of the market.

Data Wrangling and Cleaning

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. Additionally, the webpages also contain sponsored mobile advertisements and specifications for unrelated products for our purposes, such as laptops and cameras. To solve these issues, we used a for loop and extensive filtering in python to iterate over these web pages to get the necessary specifications for each mobile phone. Additional time was spent removing "rare" specifications which were not found in most mobiles. We knew that with the small number of phones, these features would either be insignificant or not statistically viable because of the rarity. So, those "rare" specifications were not included. The data wrangling process for Flipkart was like that of Gadget360. There were a few new challenges, however. String manipulation was needed to remove currency signs and "mb" from camera specifications.

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 table 1 below.

Dataset	**2023**		**2019**	
Step	**Gadgets360**	**Flipkart**	**Gadgets360**	**Flipkart**
Initial dataset size	8,695 rows x 50 columns	1,998 rows x 17 columns	1,359 rows x 21 Columns	3,114 rows x 8 Columns
Remove useless columns	8,695 rows x 6 columns	1,998 rows x 13 columns	1,359 rows x 19 Columns	3,114 rows x 4 Columns
After missing values dropped	7,800 rows	1,998 rows	1,359	2,970 rows
After duplicates dropped	4,268 rows	1,040 rows	1,359	1,079 rows
Merging	599 rows x 17 columns		387 rows x 23 columns	

Table 1. Datasets and pre-processing summary

We kept most of the columns of Flipkart and planned to add additional columns from Gadget360 because of data overlap. There were also columns that had identical information for all phones that would not be useful. 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. 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, 12 numerical columns and the dependent variable (rating). The Summary of the dataset is described in the appendix. For the 2019 version of the datasets, we followed the same process of removing leading and trailing spaces and converting the joining columns to uppercases for the initial merging steps. Similar data cleaning procedures were done in the merged datasets afterwards.

Exploratory Data Analysis

For the numerical columns, a correlation matrix was constructed between all variables.

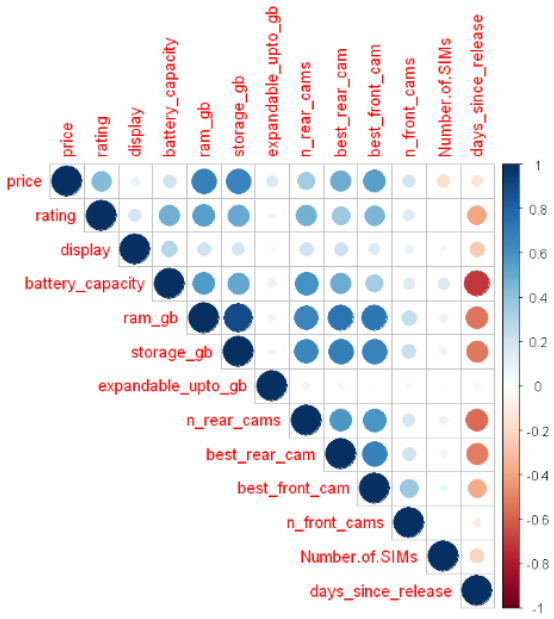


Figure 1. Correlation matrix for phone features from the merged dataset (2023)

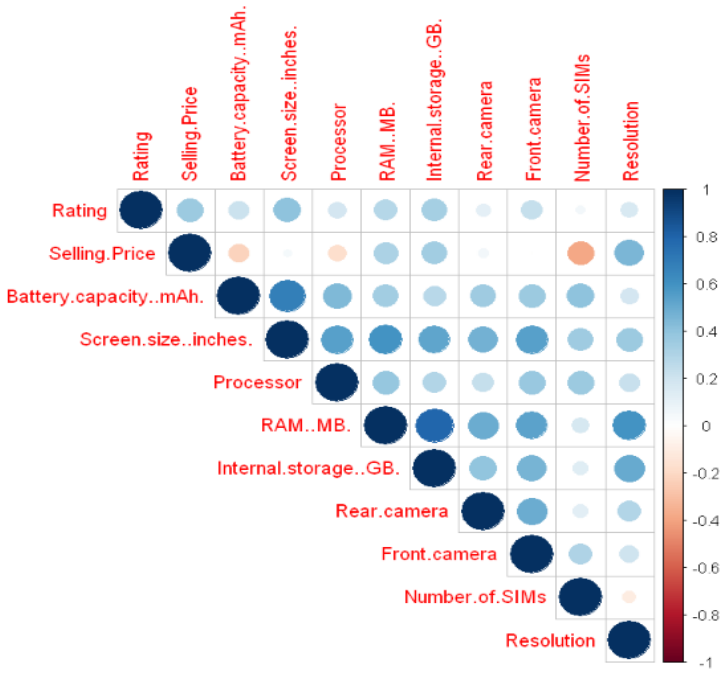


Figure 2. Correlation matrix for phone features from the merged dataset (2019)

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. (Aligns 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 distinct 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. While 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.

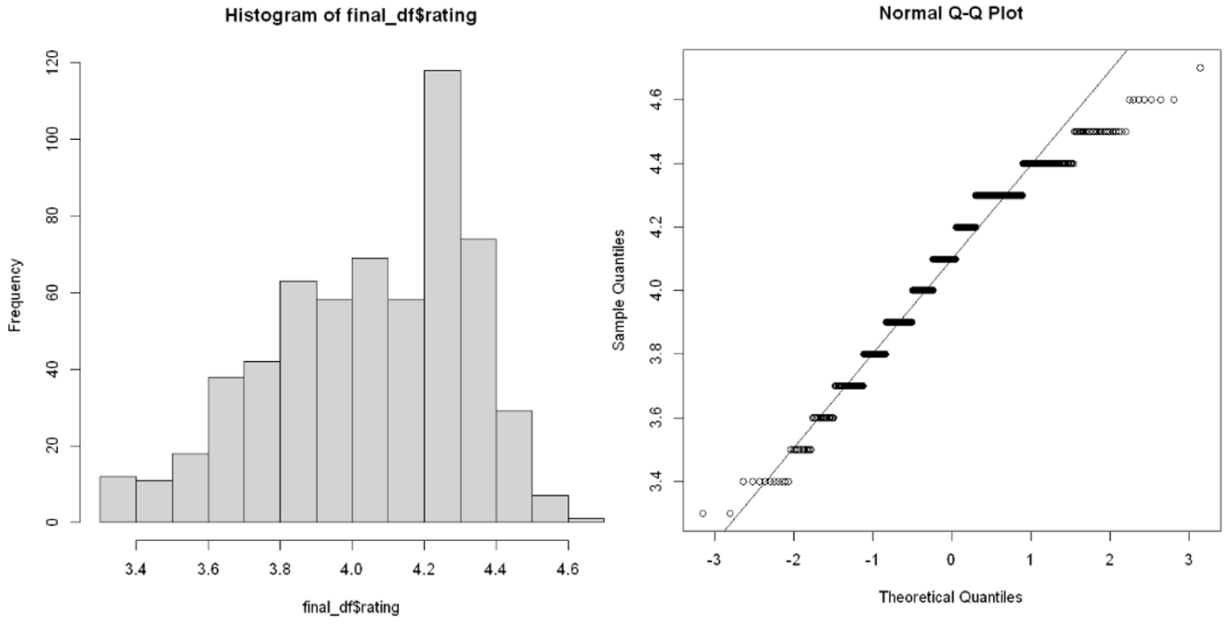


Figure 3 *(Left). Histogram of phone ratings.* ***Figure 4*** *(Right). Q-Q plot comparing data to a Normal Distribution.*

From the figure above, rating is not normally distributed and is rightly skewed. This is 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 brands that gain 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. We are also doing the same analysis on older version from 2019 to see if there are changes in what drives ratings of the four-year span.
3) **Augmentation of the Data Records of the Ecommerce Site** - The data records found in the ecommerce site Flipkart, albeit useful, can be enhanced by adding more data features 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 then joined that with the dataset of the Gadget360. Combined, there is comprehensive specification information for smartphone devices. By joining these two datasets, a more complete specification dataset is available to model their popularity.

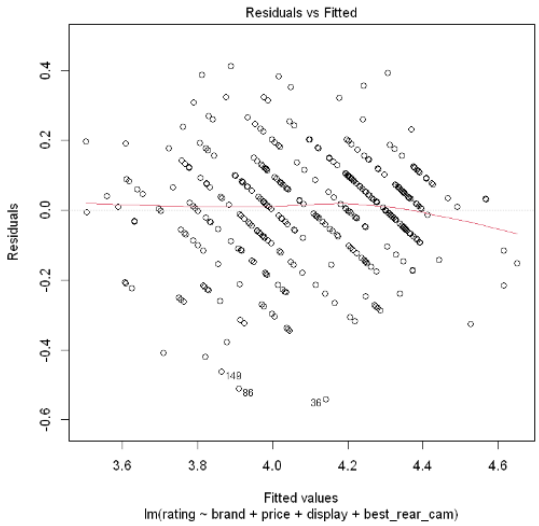
The result of our initial experiment showed that there is a high correlation between different features of smartphones. This is expected, however, as the normal convention states that increases in price will result in better specification. Based on the joined datasets from the 2019 versions, we saw that the key 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 have selected models that will fare well with multicollinearity. 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.

Methodology

The aim of our project is to predict customer phone ratings in the Indian market based on phone features. To achieve this, Multiple Linear Regression, Random Forest, and XGBoost (a distributed gradient-boosted decision tree) were used. We will begin with Multiple 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 also created a model using Random Forest. Random Forest will help to reduce the dimensionality of our data, provide a higher level of accuracy in predicting useful features, and can easily be understood. The third model we tested with, XGBoost, was selected to see if a boosting model could provide better results than a bagging model like Random Forest. Since it is trained in such a way that it can correct the error of different nodes, XGBoost might be able to capture more complex patterns in the data. 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.

Additional outliers were removed using Cook's distance and leverage points that were greater than one. The data was then partitioned into a training set with 70% of the rows, and a testing set with the remaining 30% of rows. Identical methodology was used in both the 2019 and 2023 datasets. All numbers reported below are the results of the test set.

Multiple Linear Regression – An initial model with all available predictors was created to identify significant predictors as determined by Pr(>|t|) less than 0.05. The remaining significant predictors were used in a simplified model. Those predictors were brand, battery capacity, and internal storage for 2019 and brand, price, display, and rear camera quality for 2023. The results for residuals vs fitted are available in the charts below; it was our second-best model. These models were quality checked using Q-Q, Scale-Location, and Residuals vs Leverage plots (Appendix item 4) that showed that the models were valid.



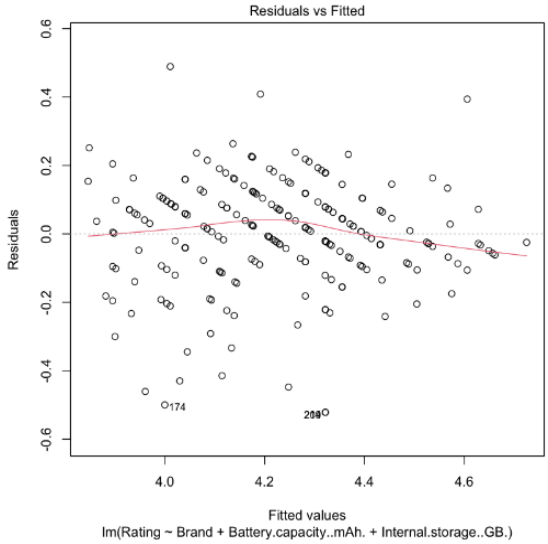


Figure 5 *(Left). 2023 model residual vs fitted plot.* ***Figure 6*** *(Right). 2019 model residual vs fitted plot.*

Random Forest - Our random forest model used 1,000 trees. Feature selection is embedded in this model, so all features were input. For both year's dataset, the best model according to the R-squared values, MSE, and RMSE was the Random Forest model. And of the two random forest models, the top eight most important features according to their node purity are listed below:

Most Important Features			
Node Purity	2019	2023	Node Purity
4.01	Brand	Brand	10.19
2.40	Price	Screen Size	4.01
1.44	Screen Size	Price	3.00
1.17	Battery Capacity	Battery Capacity	2.33
0.95	Resolution	Days Since Release	1.71
0.56	Storage	Storage	1.62
0.48/0.44	Camera Quality	RAM	1.42

Figure 7. Features Importance as defined by Random Forest models for 2019 and 2023 datasets.

XGBoost – The verbose option for XGBoost was used, as well as a max depth of 3 and a learning rate of 0.1. This boosting model had comparable results to our other models, but did have the R-squared, MSE, and RMSE of the three models used. For 2019, the most prominent features were Price, Battery Capacity, Screen Size, and Brand. For 2023, the most prominent features were Price, Screen Size, Days Since Release, and Brand. Full results are available in the appendix.

Modeling Results Summary

2023	Linear	Random Forest	XGBoost
R-Squared	66.41%	67.70%	62.98%
Mean Squared Error	0.026	0.025	0.029
Root Mean Squared Error	0.161	0.160	0.169

2019	Linear	Random Forest	XGBoost
R-Squared	57.38%	63.20%	56.10%
Mean Squared Error	0.024	0.023	0.026
Root Mean Squared Error	0.156	0.153	0.162

Table 2. Comparison of model results.

Conclusion

The results from our modeling imply that our initial hypothesis was close, but not perfect. We were correct in predicting that battery capacity and larger screens would be significant factors. However, we did not think that price and brand would trump those features. We also believed camera quality and processing speed (RAM) would be more important than the ending up being. The models indicate that brand, price, display, battery capacity, and storage are among the top five most consistently important factors. We had also alluded to the fact that the cell phone market was a rapidly changing one and that over a four-year period, the predictors of popularity would be different. This was semi-correct; the cell phone market is changing, but not as quickly as we hypothesized. The main difference between 2019 and 2023's important features was the decreased importance of the cell phone's camera quality and resolution. On the other hand, there was a minor increase in the importance of processing speed and recency of release of the cell phone. Brand, Price, Battery Capacity, and Screen size remained the most important features over this four-year span.

Potential Further Analysis and Business Impacts

While it was concluded that expensive, well-known brands are receiving the highest rating, it could be challenging for smaller brands to ever compete with the likes of Samsung or Apple. Additional analysis into the features for phones at different price points could yield different results. Consumers of low-cost phones could have different wants and needs to that have consumers with a higher budget. Keying in on the features important to low-cost or medium-cost phones could provide a business advantage for new entrants looking to avoid direct competition with top tier cell phone brands. In a future iteration of analysis, we can investigate feature importance for multiple price points or bins.

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] M. Dalal, "livemint," 10 September 2017. [Online]. Available: https://www.livemint.com/Companies/AgpQ0QQSCWBW3v3oTnFcLL/How-smartphones-changed-Flipkarts-fortunes.html.

[4] 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.

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

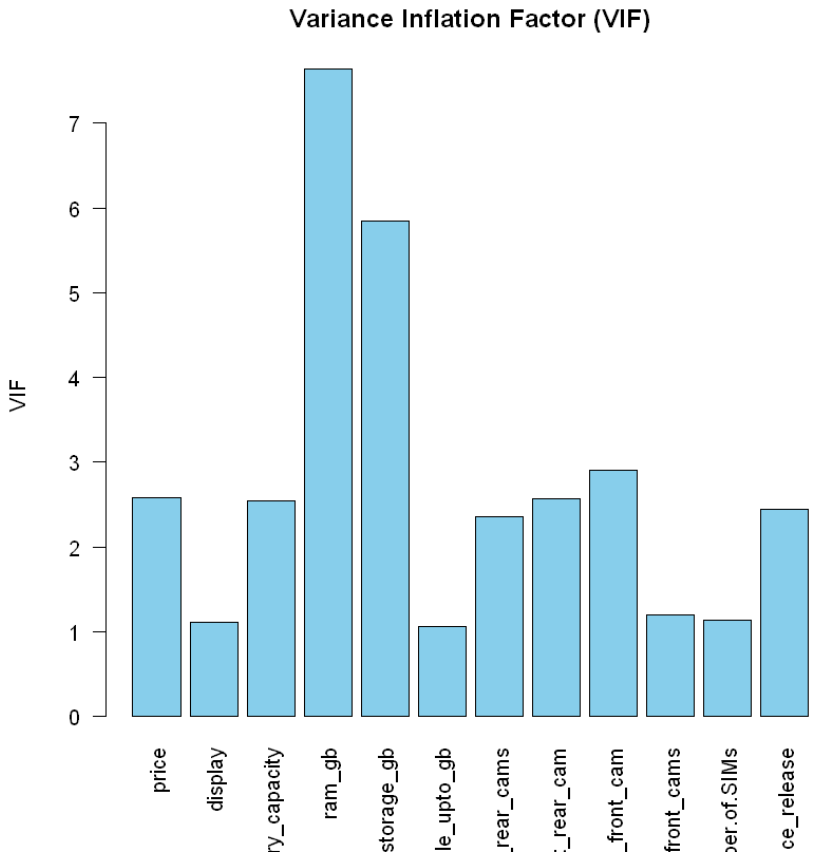
[6] Odaymat, Rommy. (2019). Factors affecting mobile phone selection. https://www.researchgate.net/publication/333461879_Factors_affecting_mobile_phone_selection

Appendix

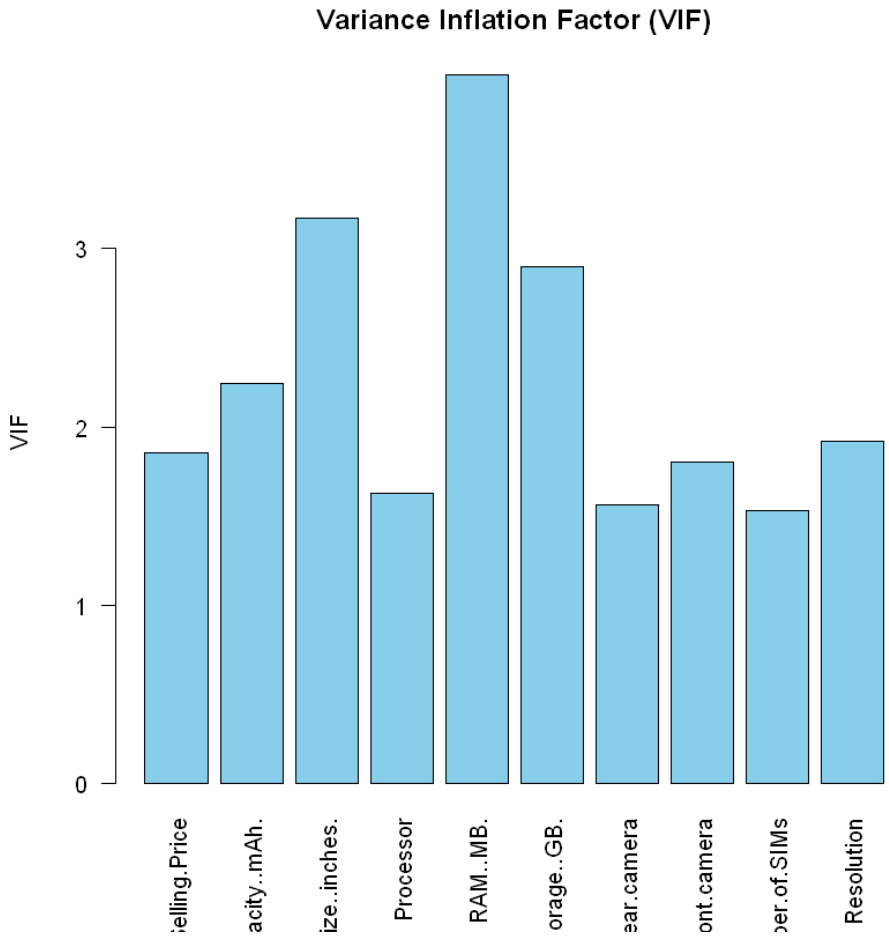
Appendix – 1: Summary of Final 2023 Dataset

```
      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
```

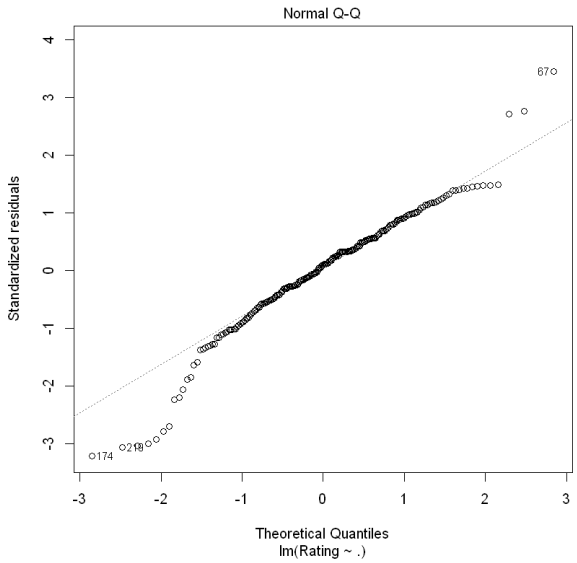
Appendix – 2: 2023 Dataset Variable Inflation Factor (VIF)

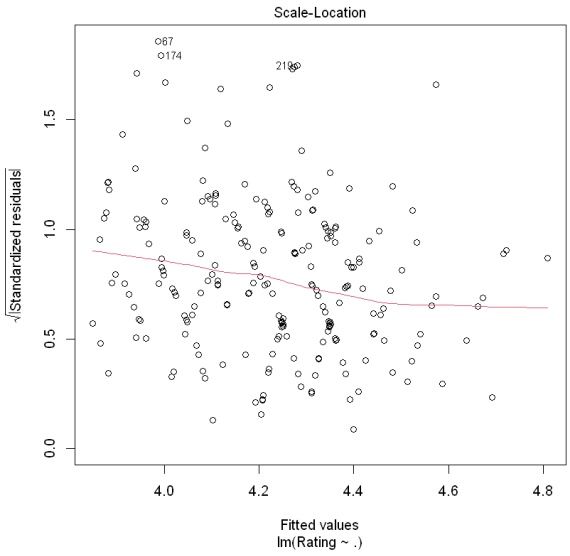


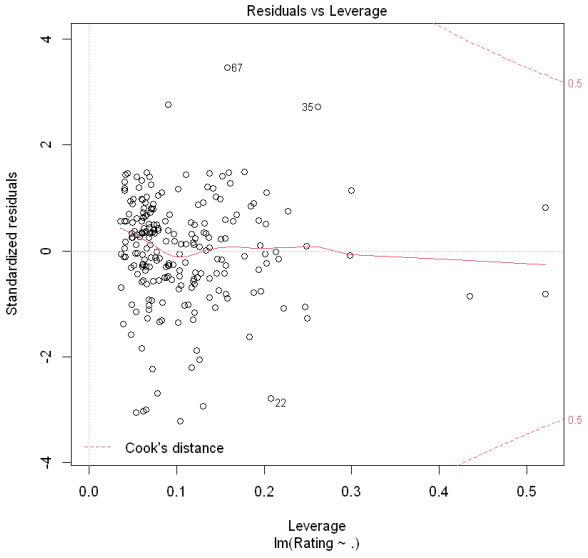
Appendix – 3: 2019 Dataset Variable Inflation Factor (VIF)



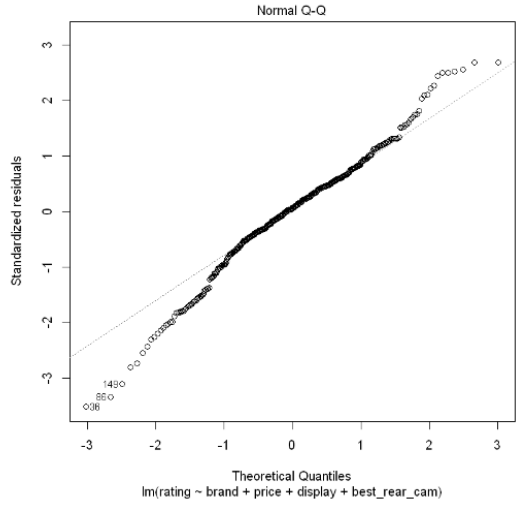
Appendix – 4.1: 2019 Multiple Linear Regression Model Diagnostic Plots

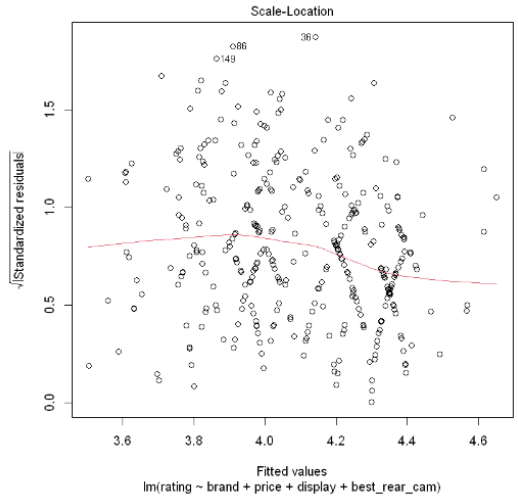


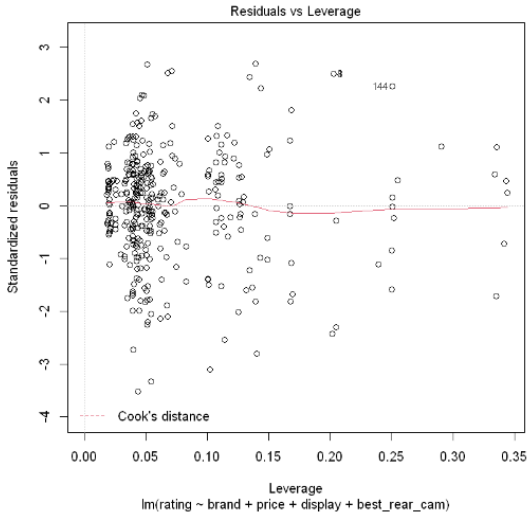




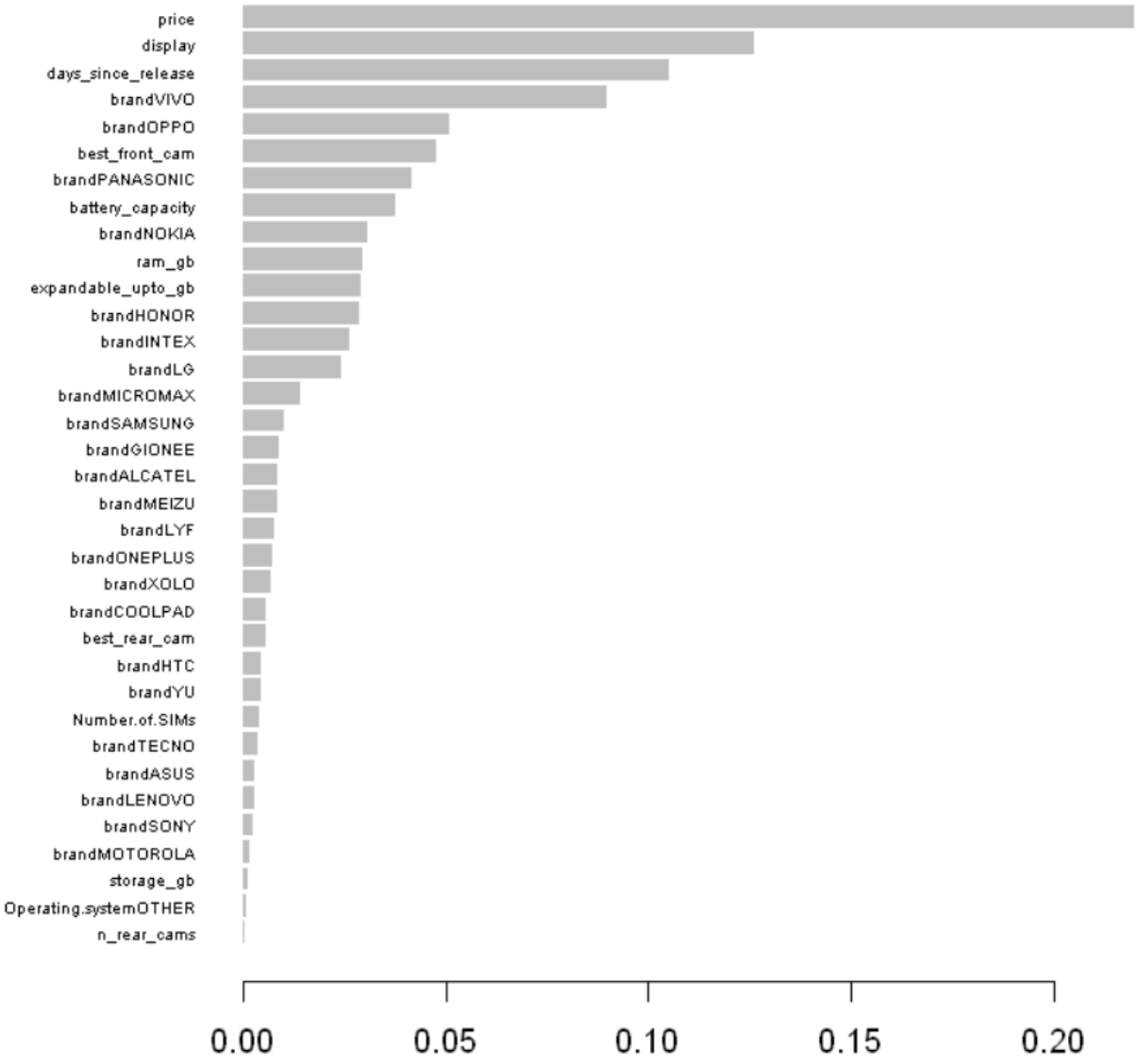
Appendix – 4.2 2023 Multiple Linear Regression Model Diagnostic Plots







Appendix – 5: XGBoost 2023 Model Variable Importance



Appendix – 6: XGBoost 2019 Model Variable Importance