**SHRUTHI BASHETTI**

**WEBSTER UNIVERSITY**

**CSDA 6010 DATA ANALYTICS PRACTICUM**

**SPRING 2024**

**Dr. OVLIA, ALI**

**Table of Contents**

[**INTRODUCTION** 4](#_Toc161819801)

[Business Problem 4](#_Toc161819802)

[Objective 4](#_Toc161819803)

[Analytical Approach 4](#_Toc161819804)

[**EXPLORATORY ANALYSIS** 5](#_Toc161819805)

[Attributes Explanation 5](#_Toc161819806)

[Data Exploration 6](#_Toc161819807)

[Data Analysis 10](#_Toc161819808)

[Predictors Analysis and Relevancy 20](#_Toc161819809)

[Data transformation 21](#_Toc161819810)

[Dimension reduction. 21](#_Toc161819811)

[Data partitioning methods 22](#_Toc161819812)

[**CLASSIFIER MODEL SELECTION** 24](#_Toc161819813)

[Models for predicting the price of used phone. 24](#_Toc161819814)

[Multi linear regression 24](#_Toc161819815)

[Decision tree for predicting price. 27](#_Toc161819816)

[Neural Network for predicting price. 29](#_Toc161819817)

[Models for Classifying the phone. 31](#_Toc161819818)

[Multinomial Logistic Regression 31](#_Toc161819819)

[Classification Tree 32](#_Toc161819820)

[Naïve Bayes: 34](#_Toc161819821)

[KNN 35](#_Toc161819822)

[Models Performance comparison For Predicting Price 37](#_Toc161819823)

[Model Evaluation for Predicting the price of a used phone using final selected model. 37](#_Toc161819824)

[Models Performance Comparison for Classification of device 38](#_Toc161819825)

[Model Evaluation for Classification of phone using final selected model. 39](#_Toc161819826)

[**CONCLUSION** 41](#_Toc161819827)

[**EXECUTIVE SUMMARY** 42](#_Toc161819828)

**Table of Figures**

[Figure 1 : structure of dataset 7](#_Toc161583327)

[Figure 2 : Head of dataset 7](#_Toc161583328)

[Figure 3: Summary Statistics of the Dataset 8](#_Toc161583329)

[Figure 4: Count of Devices 9](#_Toc161583330)

[Figure 5: Count of OS 9](#_Toc161583331)

[Figure 6: Count of 4G 9](#_Toc161583332)

[Figure 7: Count of 5G 10](#_Toc161583333)

[Figure 8: Missing value Count. 10](#_Toc161583334)

[Figure 9: Missing value count After Replacing 10](#_Toc161583335)

[Figure 10: Records with 0 values 11](#_Toc161583336)

[Figure 11: Variance of attributes. 12](#_Toc161583337)

[Figure 12: Count Plot of Device Type 12](#_Toc161583338)

[Figure 13: Bar Plot of OS 13](#_Toc161583339)

[Figure 14: Bar plot of 4G 13](#_Toc161583340)

[Figure 15: Bar Plot for 5G 14](#_Toc161583341)

[Figure 16: Bar Plot for Year released 15](#_Toc161583342)

[Figure 17: Frequency Distribution of Used Phone Price 16](#_Toc161583343)

[Figure 18: Frequency Distribution of New Phone Price 17](#_Toc161583344)

[Figure 19: Frequency Distribution of Screen Size 17](#_Toc161583345)

[Figure 20: Frequency Distribution of Internal Memory 18](#_Toc161583346)

[Figure 21: Plot for RAM 19](#_Toc161583347)

[Figure 22: Frequency Distribution of Battery 19](#_Toc161583348)

[Figure 23: Frequency Distribution of Days Used 20](#_Toc161583349)

[Figure 24: Scatter Plot of Used Price and New Price 20](#_Toc161583350)

[Figure 25: Box Plot Distribution of Screen Size 21](#_Toc161583351)

[Figure 26:Box Plot Distribution of Internal Memory 21](#_Toc161583352)

[Figure 27: Box Plot Distribution of Battery 22](#_Toc161583353)

[Figure 28: Correlation Matrix 23](#_Toc161583354)

# **INTRODUCTION**

The mobile phone market is very dynamic in today’s world, for buyers and sellers it is very important to understand the factors influencing used smartphone pricing is crucial. The pricing of a used phone can vary significantly, with an abundance of features ranging from screen size, camera quality, internal memory, ram, 4G, 5G and more. This study explores the complex relationships between the fundamental phone attributes and their impact on pricing strategies within the industry. By using predictive modeling and exploratory data analysis, we want to identify the primary factors influencing the costs of both new and used mobile phones.

Our aim is to provide business owner who will buy and sell the used phones with valuable insights by employing regression and classification models. These analytical tools will enable informed decision-making in the complex ecosystem of the mobile phone market. Through thorough data exploration and model analysis, our aim is to predict price trends and provide actionable recommendations to improve the robustness and accuracy of future models.

## Business Problem

The used smartphone market is a dynamic environment influenced by various factors, including brand, operating system, screen size, camera quality, memory, and more. However, navigating this market can be challenging due to the multitude of variables impacting pricing. For the business owner of mobile phone businesses who buy and sell used phones, understanding how these features affect pricing is crucial for making informed decisions And also to classify the phone based on the configurations.

## Objective

The primary goal of this project is to develop models that predict and classify the prices of used smartphones based on various attributes like brand, operating system, screen size, camera features, and memory. By utilizing exploratory data analysis and machine learning techniques, our objective is to provide valuable insights for sellers and buyers in the used mobile phone market. These models aim to assist stakeholders in making informed decisions regarding pricing and acquiring mobile devices, while also classifying phones into categories such as "High," "Medium," and "Low" based on their configurations which is very helpful for the business to maintain the stocks which gives them the high profits.

## Analytical Approach

The analytical goal of this project is to predict and classify used smartphone prices based on a range of attributes such as device brand, operating system, screen size, camera specifications, memory capacity, and other relevant features. The approach involves thorough data preprocessing, including attribute definition, exploration, handling missing values and outliers, assessing predictor relevance, and potentially reducing dimensionality. Subsequently, the data will be partitioned for model training and validation. Classification and regression models will be selected and fitted to the prepared dataset, with performance evaluation conducted to determine their effectiveness in predicting and classifying smartphone prices. The project aims to provide actionable insights for business owners in the mobile phone industry to optimize pricing strategies and enhance decision-making processes.

# **EXPLORATORY ANALYSIS**

## Attributes Explanation

Device\_brand: This attribute refers to the brand or manufacturer of the smartphone.

OS: The attribute OS stands for Operating System and indicates the software platform running on the smartphone, such as iOS, Android, or others.

Screen\_size: This attribute ‘screen size’ represents the diagonal measurement of the smartphone screen in inches.

4G: This attribute indicates whether the smartphone supports 4G network connectivity. It typically takes a binary value, with 1 indicating support and 0 indicating lack of support.

5G: Like the '4g' attribute, '5g' indicates whether the smartphone supports 5G network connectivity.

Rear\_camera\_mp: This attribute represents the resolution of the rear-facing camera in megapixels.

Front\_camera\_mp: This attribute represents the resolution of the front-facing (selfie) camera in megapixels.

Internal\_memory: The attribute ‘Internal memory’ refers to the storage capacity of the smartphone, typically measured in gigabytes (GB).

RAM: The attribute RAM (Random Access Memory) refers to the amount of memory available for running applications and processes on the smartphone, typically measured in gigabytes (GB).

Battery: This attribute indicates the amount of charge the smartphone's battery can hold, usually measured in milliampere-hours (mAh).

Weight: This attribute ‘Weight’ denotes the mass of the smartphone, usually measured in grams (g).

Release\_year: This attribute indicates the year when the smartphone model was released to the market.

Days\_used: This attribute represents the number of days the smartphone has been used since its purchase.

Normalized\_used\_price: This attribute refers to the price of the smartphone in a standardized format, often adjusted for factors such as depreciation and market fluctuations.

Normalized\_new\_price: This attribute represents the original price of the smartphone when it was brand new, standardized for comparison purposes.

## Data Exploration

The Given Dataset has 3454 observations of 15 attributes, of which 11 attributes are numeric attributes and 4 are non-numeric (categorical) attributes.

A screenshot of a computer

Description automatically generated

Figure 1 structure of dataset

To understand the data, head () function is used to get the first 6 rows and summary() function for the statistics of the data.

**Head of Dataset:**

**A screenshot of a computer

Description automatically generated**

Figure 2 Head of dataset

**Summary Statistics of the Dataset:**

**A group of text on a white background

Description automatically generated**

Figure 3 : Summary Statistics of the Dataset

Summary() gives mean, median, maximum, minimum, 1st quadrant, 3rd quadrant details of each numeric attribute whereas for categorical, it gives total count and its class. The dataset contains information on 3454 observations, with categorical variables like device brand and operating system. Screen sizes range from a minimum of 5.08 to a maximum of 30.71 units, providing a diverse spectrum of display dimensions among the phones. Categorical variables X4g and X5g denote whether a phone supports 4G and 5G capabilities, respectively. Camera specifications reveal that the mean rear camera megapixels is 9.46, and the mean front camera megapixels is 6.554. Internal memory and RAM, representing storage and processing capabilities, have mean values of 54.57 and 4.036, respectively. Battery capacities vary, with a minimum of 500 and a maximum of 9720, yielding a mean of 3133. Phone weight ranges from 69.0 to 855.0 units, with an average of 182.8. Release years span from 2013 to 2020, offering a temporal context for the dataset. The "days\_used" variable indicates usage patterns, with a mean of 674.9. Normalized used and new prices provide standardized metrics for price comparison, with mean values indicating central tendencies.

Notably, certain variables such as rear\_camera\_mp, front\_camera\_mp, internal\_memory, ram, battery, and weight have missing values, as indicated by the NA entries in the summary.

**A group of text on a white background

Description automatically generated**

Figure 4 : Count of Devices

From fig 4, we can see all the device\_brands and the respective no of records for each brand. There are 502 records with no brand name and referred as ‘Others’ followed by Samsung with 341 records and lowest in ‘Infinix’

**A group of black text

Description automatically generated**

Figure 5: Count of OS

From fig 5, we can see that there are four types of operating system offered and most of the smartphones are using ‘Android’ with 3214 records and the least is ‘iOS’ with 36 records.

A white background with black text

Description automatically generated

Figure 6: Count of 4G

A white background with black text

Description automatically generated

Figure 7: Count of 5G

From the fig 6, fig 7 the attribute 4g has two types of values i.e yes and no with 1119 and 2335 observations. 5g with no as 3302, yes 152 observations.

**Handling NA’s:**

From below snip, we can see that rear\_camera, front\_camera, internal memory, ram, battery, weight has na values in them.

A white background with black text

Description automatically generated

Figure 8: Missing value Count.

These were handled by replacing the missing values with the mean value, within each group defined by the combination of "device\_brand" and "os." Surprisingly, we still had more na values as shown below fig 9.

A close-up of a screen

Description automatically generated

Figure 9: Missing value count After Replacing

In further investigation, understood that all the 10 missing values are from Infinix brand phone with Android as os. Here, its evident that these phones having front camera which suppose to have back camera as well. Decided to replace these na values with 0 instead of removing them completely. After replacing na values of Infinix brand phone with android on os, there are no more NA’s left.

**Handling 0’s:**

There are 30 records having 0’s in front camera attribute. But all these records are from Nokia phone with others as OS. Since it could be a old model phone, assuming that these phones did not have front camera.

A screenshot of a computer screen

Description automatically generated

Figure 10: Records with 0 values

**Handling Data Types:**

Based on the models we apply for regression and classification, it will be decided later in order to convert device brand, OS into dummies or not.

## Data Analysis

As part of analysis, variance for numeric attributes are as shown in fig 11. Variance values provide essential insights into the dispersion and influence of features, High variances indicate greater diversity, while low variances suggest more consistent attributes. All these attributes are having good variance.

A close-up of a computer screen

Description automatically generated

Figure 11: Variance of attributes.

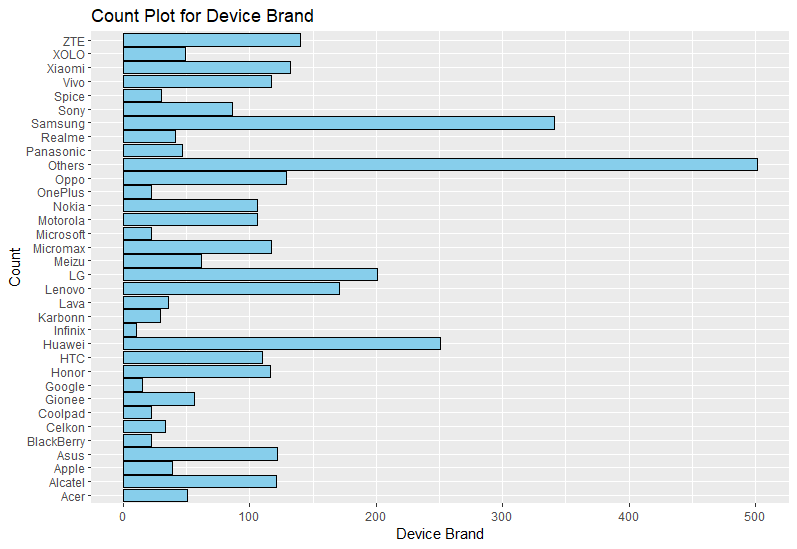


Figure 12: Count Plot of Device Type

The count plot fig 12, illustrates the distribution of devices in the dataset, highlighting that the category "others" contains a larger number of phones and the Samsung phones. We have 10 Infinix phones which are the least number of phones in the dataset.

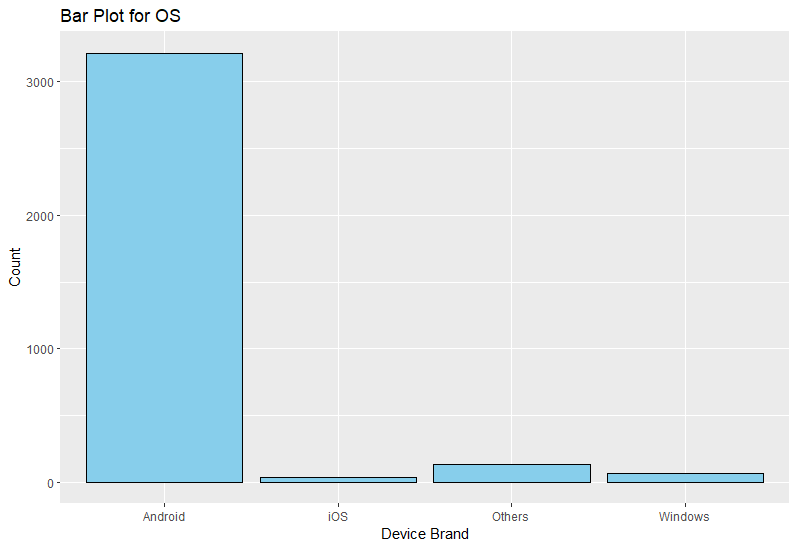


Figure 13: Bar Plot of OS

The bar plot fig 13, shows that we have 4 types of OS in our system with more number of phones having android as OS.

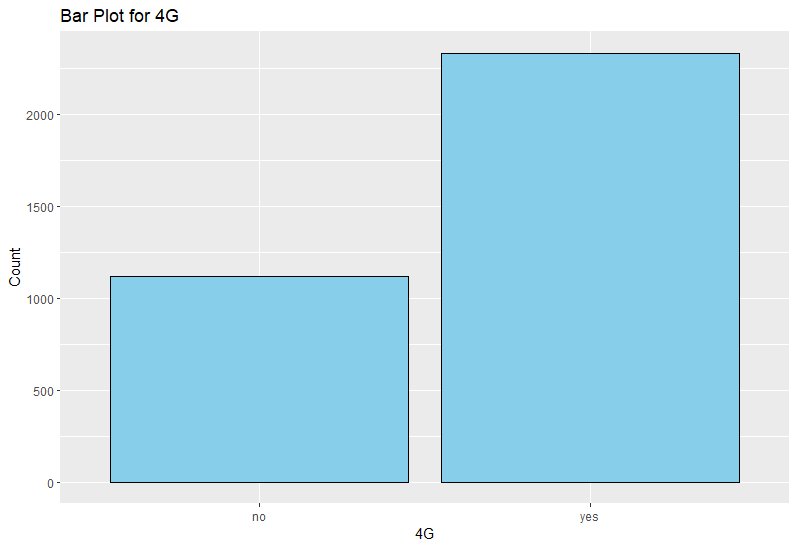


Figure 14: Bar plot of 4G

The bar plot in fig 14, shows the distribution of 4g phone in the dataset, more than 2500 out of 3450 phones support 4g connectivity.

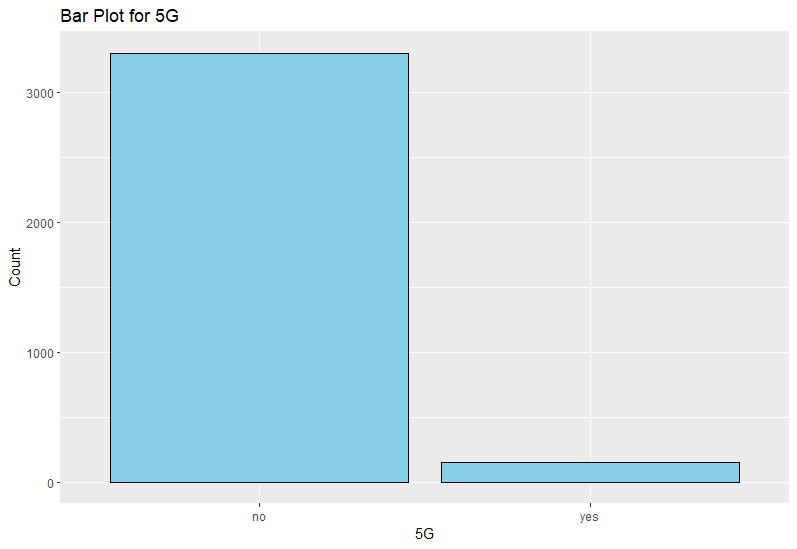


Figure 15: Bar Plot for 5G

The above bar plot, fig 15 shows the distribution of 5g phone in the dataset, very nominal count of phones in the dataset support 5g connectivity.

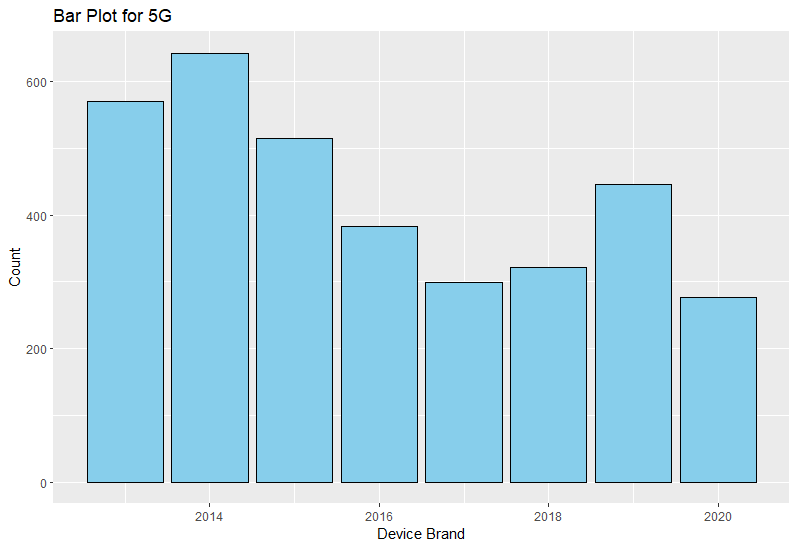


Figure 16: Bar Plot for Year released

The bar chart in fig 16 shows, more number of phones that we have in the dataset are 2014 model phones and then 2013 model phones.

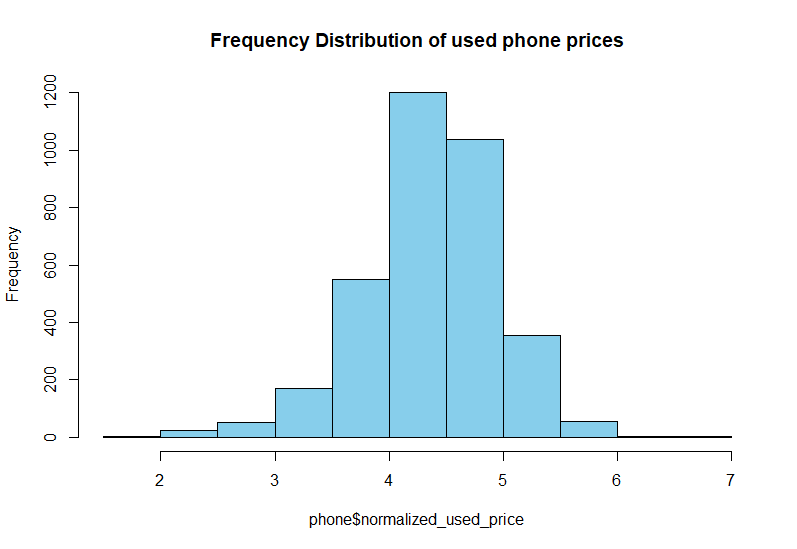
From the above, we can say that most releases are in the year 2014 followed by 2013 and then 2015. The year 2020 holds the lowest releases of phones. The phones in the dataset are models between 2013-2020.

Figure 17: Frequency Distribution of Used Phone Price

The histogram in fig 17 shows the distribution of used prices we have, there are more phones sold at lower prices than at higher prices. The most common price for a used phone appears to be between 2 and 5.

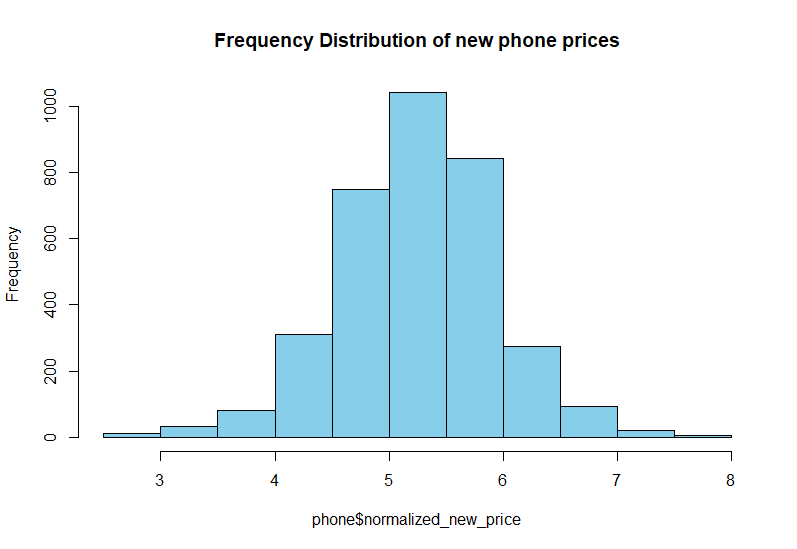


Figure 18: Frequency Distribution of New Phone Price

The histogram in fig 18 shows the distribution of prices of new phones which is in normal distribution since the prices are already normalized. More number of phone having prices between 4.5 – 6

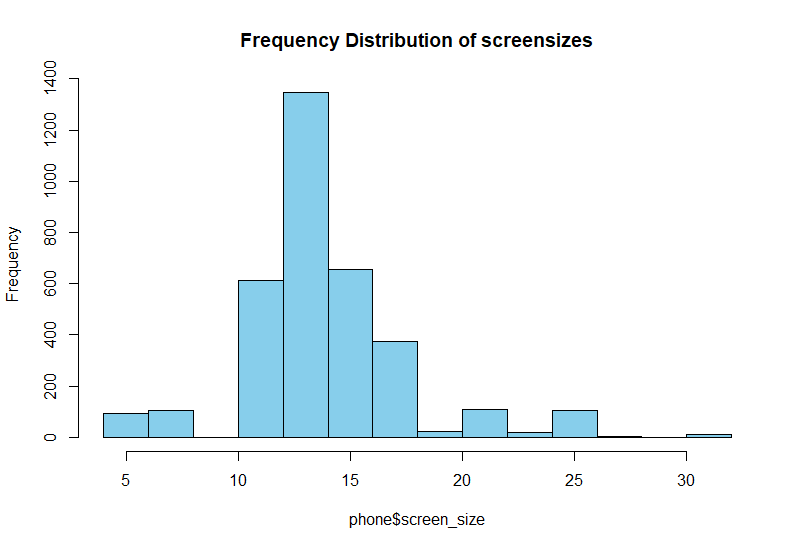


Figure 19: Frequency Distribution of Screen Size

The above histogram in fig 19 shows the distribution of screen sizes we have in the dataset, it is right skewed. There are more phones between 10-15 as a screen size. The right corner are phones with size above 30 can be considered as Tablets.

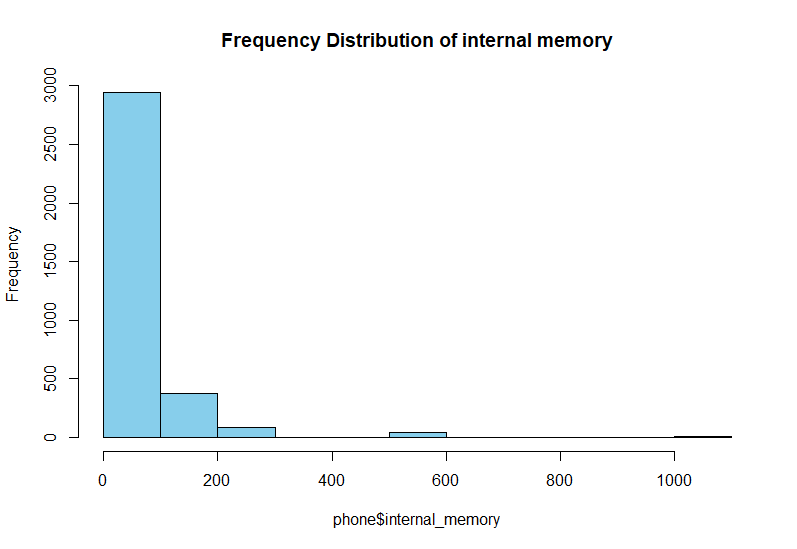


Figure 20: Frequency Distribution of Internal Memory

The histogram in fig 20 shows the distribution of internal memory, around 3000 phones having internal memory between 0-100 and few phones with internal memory of 126, and 256.

There are very few phones with internal memory of 1024 GB on right corner of the graph which are not outliers.

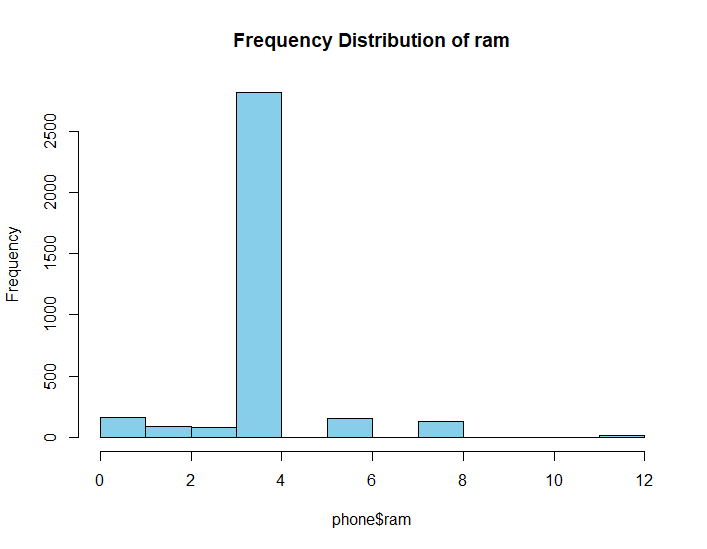


Figure 21: Plot for RAM

The plot fig 21 shows the distribution of RAM, most of the phones are with 4GB of RAM. There are phones with 1 GB of RAM, 12 GB of RAM which are not outliers, but the phones can have 1GB of RAM as well as 12 GB of RAM.

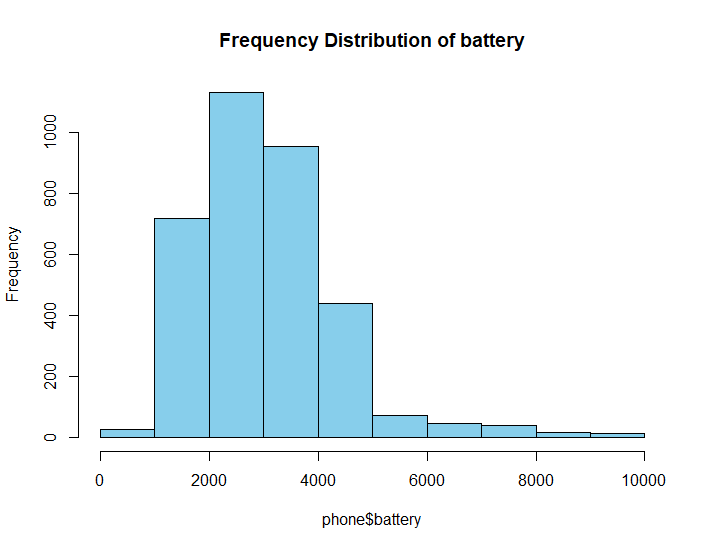


Figure 22: Frequency Distribution of Battery

The above plot in fig 22 shows the distribution of Battery, where most of the phones having battery between 1000-4000 and good amount of phones with battery between 4000 to 5000 mah.

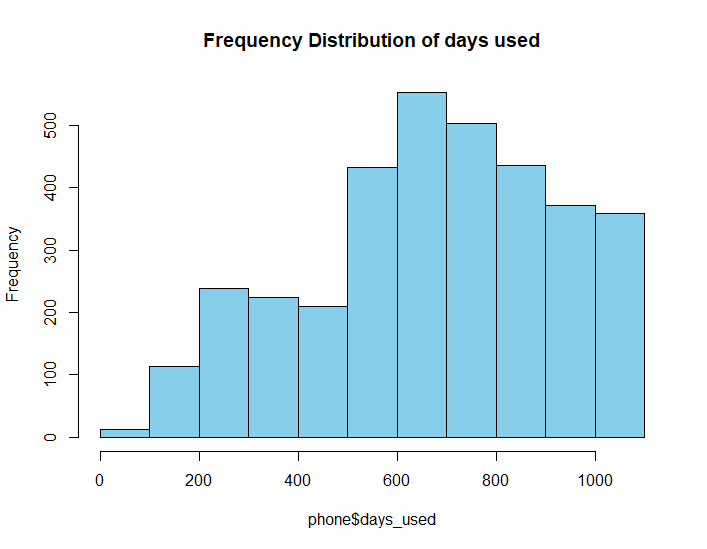


Figure 23: Frequency Distribution of Days Used

This plot in fig 23 shows the distribution of number of days used and phones were used between 1 year to 3 years.

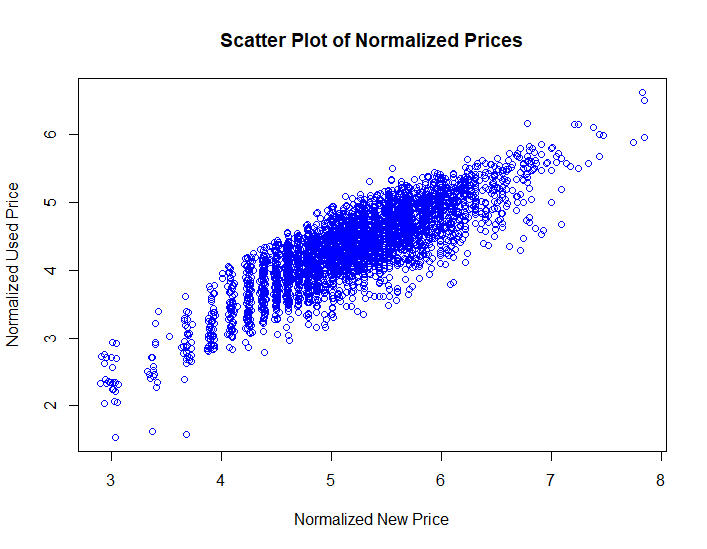


Figure 24: Scatter Plot of Used Price and New Price

The scatterplot in fig 24 shows the correlation between used price and new price which are actually in good correlation.

We can build reliable models when we have predictors without outliers. I have used boxplots with respect to target to see the outliers.

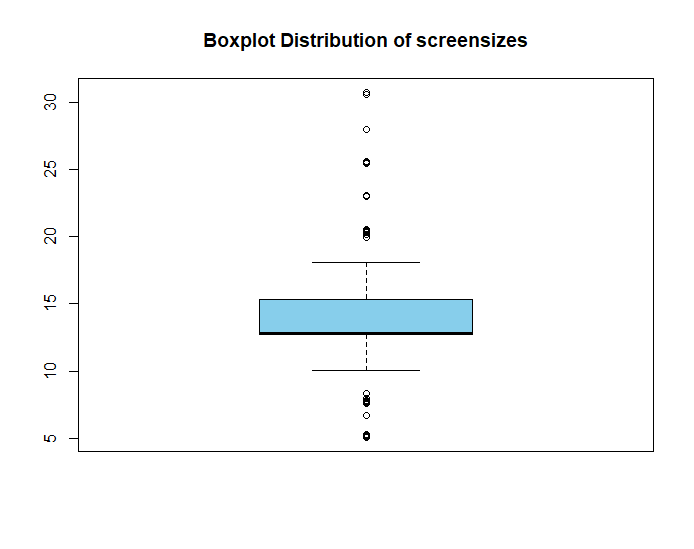


Figure 25: Box Plot Distribution of Screen Size

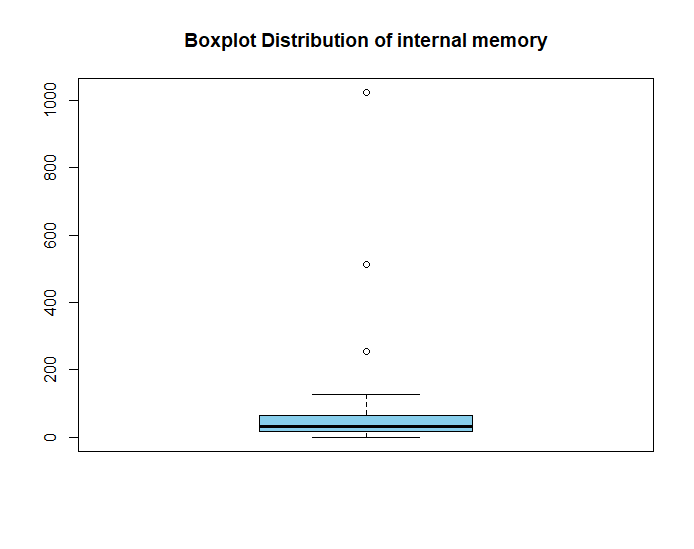


Figure 26:Box Plot Distribution of Internal Memory

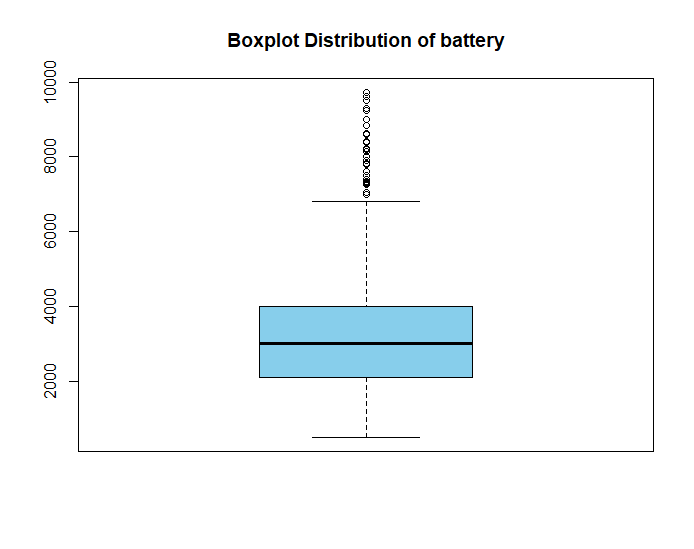


Figure 27: Box plot Distribution of Battery

Though the boxplot shows outliers in the dataset, those should not be considered as outliers since we can have phones with 25-30 inch as well as phones with 5-10 inch of screen size. Same with internal memory and battery, we can have phones with 1024 GB of internal memory, battery with 10000 MAH. so, these are all not outliers.

## Predictors Analysis and Relevancy

As part of the predictor analysis and relevancy, correlation coefficients were calculated among the numeric attributes in the dataset. This involved assessing the strength and direction of linear relationships between pairs of variables. The correlation results offer insights into potential associations and dependencies, aiding in the identification of influential predictors and potential multicollinearity.

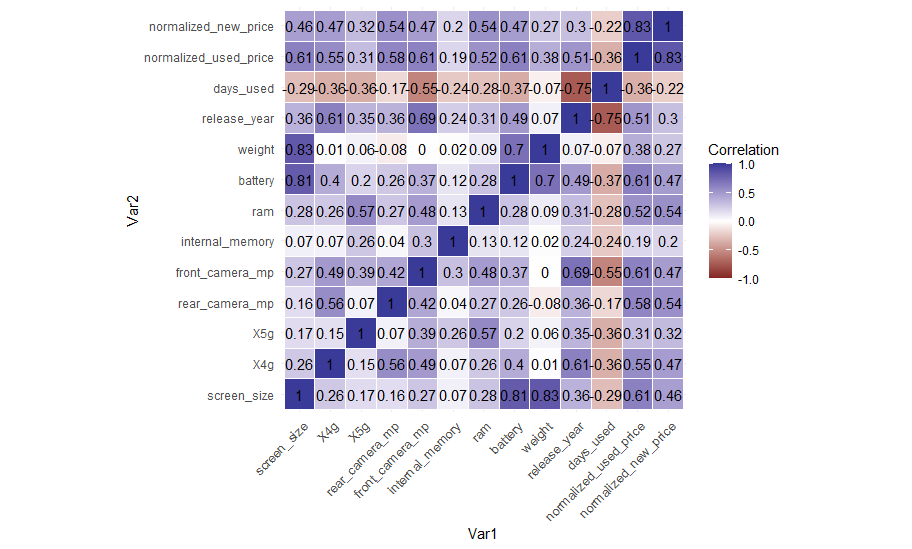


Figure 28: Correlation Matrix

The attributes like screen size, X4g, rear\_camera\_mp, front \_camera\_mp, RAM, Battery, release year, normalized new price are in strongly correlation with used price of a phone. The attributes X5g, internal memory, weight are not in correlation with used prices of phone.

Apart from the numeric attributes, we have 2 categorical attributes which are device brand, OS which need to be considered for prediction of used phone’s prices.

## Data transformation

As part of data transformation and feature engineering, added new attribute named ‘class’ in order to classify the phone’s quality or configuration based on certain attributes like screen size, RAM, Internal memory and battery.

These attributes such as RAM Internal Memory, Screen Size, Battery are categorized into classes (Low, Medium, High) based on predefined class ranges. All the class ranges are defined based on the data i.e. if the data range between 0-12, this is split into three portions i.e. each class with range of 4 like 0-4, 5-8, 9-12.

Furthermore, a Weighted Average feature is computed by assigning weights to each class equally and calculating the weighted average of the attributes. This weighted average is then used to determine the final class of each phone, with thresholds set to classify phones as Low, Medium, or High.

The weighted Average is calculated based on the threshold defined, the threshold is defined based on the number of observations and the threshold is tweaked to balance the observations under each class.

Finally, the "class" attribute derived which says the quality/configurations of the phone based on above mentioned attributes. Using this we can build a model to classify the phone into high, medium or low quality/configured phone.

Overall, these data transformations and feature engineering steps contribute to creating a more informative and structured dataset suitable for building predictive models to estimate used phone prices and classify phones based on their attributes.

The dummy creation is essential for handling categorical variables, such as 'os' (operating system) and 'device\_brand' (brand of the device), in predictive modeling tasks.

By utilizing dummy encoding, the original categorical variables are effectively transformed into a format that machine learning algorithms can readily interpret and utilize for predictive modeling. This transformation enables the inclusion of categorical variables in predictive models, thereby improving the model's accuracy and effectiveness in capturing underlying patterns within the data. Additionally, removing the original categorical columns after dummy encoding helps to streamline the dataset and reduce redundancy, making it more suitable for subsequent analysis and modeling tasks.

## Dimension reduction.

As we have less number of predictors in the dataset, keeping the interpretability and computational efficiency of the model into consideration, chose feature selection instead of dimension reduction.

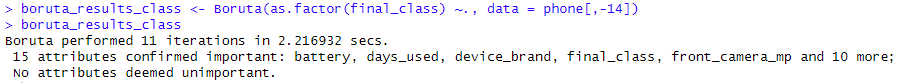
The Boruta feature selection algorithm was applied to the dataset, aiming to identify the most important attributes for predicting the normalized used price of smartphones. The algorithm iteratively evaluates the relevance of each attribute in relation to the target variable, considering interactions and dependencies among features.

A close-up of a computer code

Description automatically generated

From the output of Boruta, it's evident that after 11 iterations, all 14 attributes were confirmed as important for predicting the normalized used price.

Boruta suggests no attributes were deemed unimportant, suggesting that all features in the dataset contribute in some capacity to the prediction task. This underscores the complexity of smartphone pricing prediction and the interconnected nature of various smartphone attributes.



The Boruta feature selection algorithm was employed to reduce the dimensionality of the dataset for predictive modeling of the 'class' attribute. After 11 iterations, Boruta identified all attributes as being significantly important for predicting the target variable.

Interestingly, Boruta did not flag any attributes as unimportant, suggesting that all features in the dataset contribute to the predictive power of the model. This outcome implies that each attribute, including those not explicitly mentioned, plays a role in determining the class of the smartphone in the context of the analysis.

## Data partitioning methods

The data should be partitioned into training and validation sets and sometimes holdouts because, the training set will be used to build different models and validation to evaluate the developed models the holdout set will be used to test how well selected model performs. In this way we get an unbiased estimate of how well the model performs.

We can partition data in multiple ways based on the purpose. There are methods such as train-test, stratified sampling, k-fold cross validation, validation holdouts.

For this data, we are using train- validation-holdouts method in which we split the data into 3 subsets.

**Train dataset:**

We use train sets to build different models. The purpose is to enable models to learn from the data by capturing patterns and relationships. In this project 50% data split into a training set which is 1727 records out of 3454.

A screen shot of a computer code

Description automatically generated



Validation dataset:

The portion of data that is used to assess how well the model fits, to adjust the models, and to select the best model from models that built. It helps fine-tuning of model parameters and helps prevent overfitting. Here 25% data split into validation set which is 864 records out of 3454.



**Holdout dataset:**

The portion of data that is used only at the end of model building and selection process to assess how well the final model would perform on new or unseen data. In this project, the business owner wants 25% data split into a holdout set which is 863 records out of 3454.



We set the seed to produce the same results again.

# **CLASSIFIER MODEL SELECTION**

In predicting used mobile prices, linear regression offers a straightforward approach by modeling the relationship between predictors and price. Decision trees provide a flexible method, capturing nonlinear relationships and interactions among features. Meanwhile, neural networks excel in capturing intricate patterns and nuances in the data, offering high predictive accuracy.

For predicting used mobile class, we're employing classification tree, Naive Bayes, and neural network models. These methods offer diverse approaches to categorizing mobile classes based on their attributes, enabling accurate classification and informed decision-making.

The first model is built to predict the price of the used phone customer needs based on the specifications and the second model is to classify the phone into high-medium-low configuration categories based on the specifications of the phone which is useful for the business to convince to customer to buy the phone and in many aspects.

## Models for predicting the price of used phone.

### Multi linear regression

Multilinear regression is a statistical method used to model the relationship between multiple independent variables and a continuous dependent variable. It aims to identify the linear equation that best represents the observed data points, facilitating the prediction of the dependent variable based on the values of the independent variables. Here we build a model to predict the price based on the all predictors.

A screenshot of a computer

Description automatically generated

In order to improve the performance of the model, a stepwise method is applied on the model to know the important features as shown in below fig.

A screenshot of a computer

Description automatically generated

**Multi linear regression with selected features using stepwise**

Multi linear regression with selected features is built to improve the model’s performance.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

The model exhibits a residual standard error of 0.231, indicating the average distance between observed and predicted values. The high Multiple R-squared value (0.8422) suggests that 84.22% of the variability in the dependent variable is explained by the model. The significant F-statistic (p-value < 2.2e-16) supports the overall effectiveness of the regression model.

**Evaluation:**

A screenshot of a computer code

Description automatically generated

The summary statistics for the predicted normalized used prices show a range from 2.667 to 6.137, with a mean of 4.360. These closely align with the summary statistics of the actual normalized used prices in the validation set, ranging from 2.268 to 6.619, with a mean of 4.361. The high correlation coefficient of 0.92 indicates a strong positive linear relationship between the predicted and actual normalized used prices.



The regression model for predicted normalized used prices exhibits a slight underestimation bias (ME: -0.00154) with a low overall prediction error (RMSE: 0.2324). The model's accuracy is further evidenced by a modest Mean Absolute Error (MAE: 0.1818) and a Mean Absolute Percentage Error (MAPE: 4.33%), suggesting effective forecasting performance on the validation set.

### Decision tree for predicting price.

A decision tree is a predictive modeling tool that recursively splits data into subsets based on the most significant features, resulting in a tree-like structure of decision nodes. It's widely used for classification and regression tasks due to its interpretability and ability to handle both numerical and categorical data.

**Model:**

**A black and white text

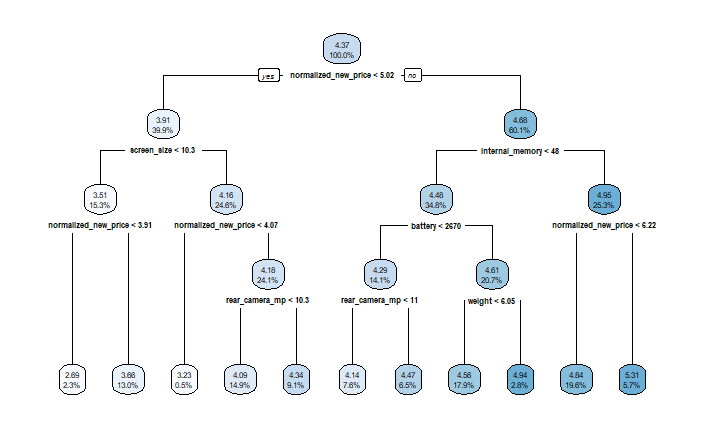
Description automatically generated**

A screenshot of a computer code

Description automatically generated

The regression tree, built for predicting 'used\_price,' initiates with a root node representing 1727 observations. It branches on features such as 'normalized\_new\_price,' 'screen\_size,' 'internal\_memory,' 'battery,' and 'rear\_camera\_mp.' Terminal nodes provide specific 'used\_price' predictions, with conditions like 'normalized\_new\_price' influencing the splits. This interpretable model captures nuanced relationships among features, offering distinct predictions in its terminal nodes and providing insights into the influential factors shaping the predicted 'used\_price.'

**Tree:**

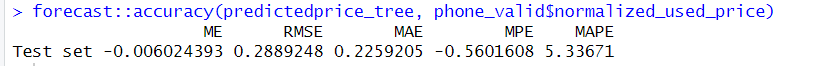


**Evaluation:**

**A screenshot of a computer screen

Description automatically generated**

The summary of the predicted 'used\_price' values from the regression tree model on the validation set reveals a range from 2.692 to 5.312, with a mean of 4.367. This closely aligns with the summary statistics of the actual 'normalized\_used\_price' in the validation set, ranging from 2.268 to 6.619, with a mean of 4.361. The high correlation coefficient of 0.8720 indicates a strong positive linear relationship between the predicted and actual 'normalized\_used\_price,' suggesting that the model captures significant patterns in the validation data.



The evaluation of the regression tree's predictive performance on the validation set indicates promising results. The predicted 'used\_price' values exhibit a mean error of -0.0062, suggesting a minor deviation from the actual values. The root mean squared error (RMSE) of 0.2889 and mean absolute error (MAE) of 0.2259 highlight the model's overall accuracy, capturing the dispersion and absolute differences between predictions and actual values. The mean percentage error (MPE) of -0.5601% and mean absolute percentage error (MAPE) of 5.3367% provide insights into the percentage-wise accuracy, indicating a generally close fit of the regression tree's predictions to the observed 'normalized\_used\_price' in the validation dataset.

### Neural Network for predicting price.

Neural networks excel in capturing intricate patterns and nuances in the data, offering high predictive accuracy.

**Model:**

A computer screen shot of text

Description automatically generated

**Plot:**

A diagram of a computer data

Description automatically generated with medium confidence

**Evaluation:**

A close-up of words

Description automatically generated



The correlation coefficient between the predicted values (unnorm\_phoneprice) and the actual 'normalized\_used\_price' in the validation set is 0.9143. This high positive correlation indicates a strong linear relationship between the predicted and observed values, suggesting that the model's predictions align closely with the actual 'normalized\_used\_price' in the validation dataset.



The assessment of forecast accuracy for the model predicting 'normalized\_used\_price' on the validation set reveals favorable results. The mean error (ME) is minimal at 0.68, indicating a slight overestimation on average. The root mean squared error (RMSE) and mean absolute error (MAE) are 0.7246 and 0.687, respectively, emphasizing the accuracy in capturing both the dispersion and absolute differences between predicted and actual values. The mean percentage error (MPE) of 13.7% and mean absolute percentage error (MAPE) of 13.76% provide insights into the percentage-wise accuracy, indicating a generally close fit of the model's predictions to the observed 'normalized\_used\_price' in the validation dataset. Overall, these metrics suggest a robust performance of the model in forecasting 'normalized\_used\_price.'

## Models for Classifying the phone.

The second model is to classify the phone into high-medium-low configuration categories based on the specifications of the phone which is useful for the business in many aspects.

### Multinomial Logistic Regression

Multinomial Logistic Regression, often abbreviated as multinom or MNL, is a statistical model employed in predictive modeling when the outcome variable has more than two categories. It extends the principles of binary logistic regression to handle multiple discrete categories. In multinomial logistic regression, the model estimates the probability of each category, enabling the analysis and prediction of outcomes across a range of distinct possibilities.

**Model:**

A close-up of a text

Description automatically generated

**Evaluation**

A screenshot of a computer

Description automatically generated

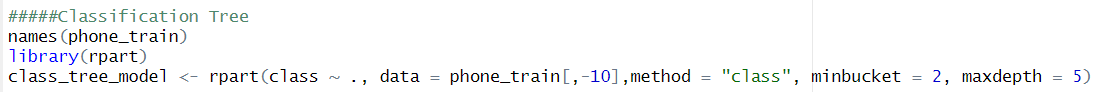
The confusion matrix and associated statistics summarize the performance of a classification model. In this example, the model predicts three classes (High, Low, Medium) against actual class labels. The overall accuracy is 33.72%, with class-specific sensitivities, specificities, positive and negative predictive values, and prevalence rates. The Kappa statistic suggests limited agreement beyond chance, and the Balanced Accuracy provides an average accuracy across classes, highlighting potential imbalances in prediction performance.

### Classification Tree

A classification tree works by recursively splitting the dataset based on the most informative features, creating a tree structure where leaves represent distinct class labels. At each node, the algorithm selects the attribute that maximizes information gain or minimizes Gini impurity. This process continues until a stopping criterion is met, resulting in a hierarchical model that efficiently classifies instances based on their feature values.

**Model:**

Built a classification tree using rpart as shown below.



A computer code with numbers and letters

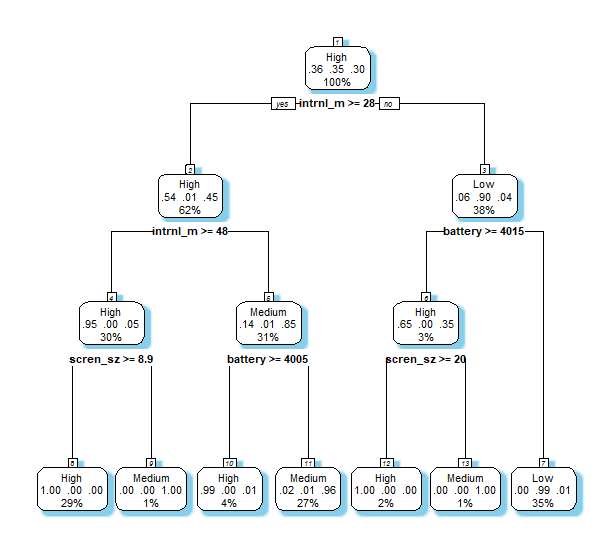
Description automatically generated

The root node represents the entire dataset of 1728 instances, with a distribution of classes (High, Medium, Low) in proportions 35.5%, 34.78%, and 29.6%, respectively and the tree branches based on conditions involving features such as 'internal\_memory,' 'screen\_size,' and 'battery’, terminal nodes marked with '\*' represent the final predicted classes for specific subsets of the data.

For instance, if 'internal\_memory' is greater than or equal to 28, the model further splits based on 'internal\_memory' and 'battery,' leading to specific predicted classes (High, Medium) in the terminal nodes.

The tree serves as a predictive model for classifying instances into 'High,' 'Medium,' or 'Low' based on the specified conditions, capturing complex relationships among the features and providing different predicted classes in its terminal nodes.

**Tree:**



**Evaluation:**

A screenshot of a computer

Description automatically generated

The classification tree model attains remarkable performance with an overall accuracy of 98.26%, showcasing its ability to effectively categorize instances into the High, Medium, or Low classes. The model's precision is further highlighted by class-specific metrics, particularly high sensitivity ranging from 97.56% to 99.17%, indicating its proficiency in correctly identifying instances within each class. Additionally, the model demonstrates strong specificity, exceeding 98.97% for the Medium and Low classes. The Kappa statistic, registering at 0.9739, underscores substantial agreement beyond chance, affirming the robustness of the classification tree in capturing intricate patterns and relationships within the dataset. This comprehensive evaluation underscores the model's reliability and effectiveness in making accurate predictions across multiple classes.

### Naïve Bayes:

Naive Bayes, a probabilistic algorithm, simplifies classification by assuming feature independence, making it efficient for various tasks such as spam detection and document categorization.

**Model:**

**A computer code with text

Description automatically generated with medium confidence Evaluation:**

A screenshot of a computer screen

Description automatically generated

The confusion matrix and associated statistics offer a comprehensive evaluation of the classification model. With an overall accuracy of 66.71%, the model outperforms the No Information Rate significantly, demonstrating its effectiveness in categorizing instances into High, Medium, or Low classes. The Kappa statistic of 0.4924 indicates moderate agreement beyond chance. Class-specific metrics reveal varying sensitivities and specificities, with notable positive predictive values for the Low class.

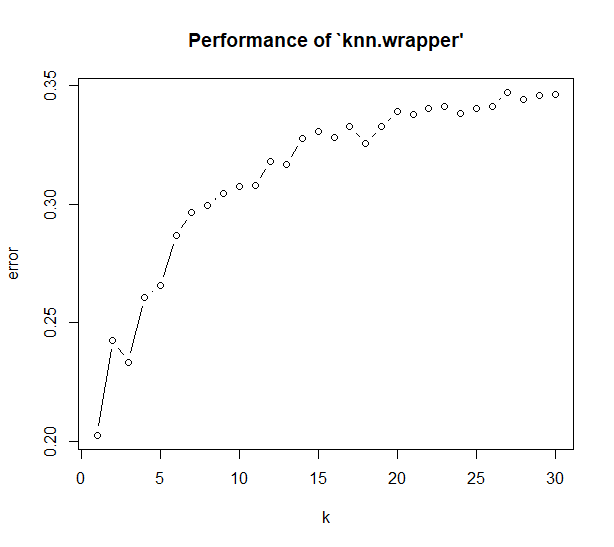
### KNN

K-Nearest Neighbors (KNN) is a simple yet effective algorithm for classification and regression tasks. It works by identifying the K nearest data points to a given observation and classifying or predicting based on the majority vote or average of their labels or values, respectively.

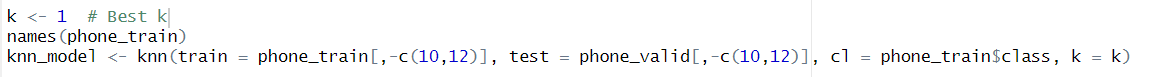
Before building a model using KNN, used tunning to get the best k value which gives the best predicting power.

A close up of text

Description automatically generated



**Model:**

****

**Evaluation:**

A screenshot of a computer

Description automatically generated

The confusion matrix summarizes the performance of a classification model. With an overall accuracy of 79.16%, the model demonstrates strong predictive ability. Class-specific metrics reveal high sensitivities (recall) for High (89.02%) and Low (81.25%) classes, while balanced accuracy reflects a good overall balance between sensitivity and specificity across classes (92.27%, 75.85%, 83.97%). The Kappa statistic indicates substantial agreement beyond chance (0.6862).

## Models Performance comparison For Predicting Price

Here we performed 3 different models for predicting the price of a used phone which are multilinear regression, decision tree and Neural Networks. All the models have their own strengths and limitations. Multilinear regression excels in capturing linear relationships and offering interpretability, while decision trees uncover complex decision rules. Neural Networks, with their capacity for learning intricate patterns, may handle non-linear relationships effectively. However, each model comes with its own set of strengths and limitations. For instance, multilinear regression assumes linearity, decision trees can be prone to overfitting, and Neural Networks may require substantial data and computational resources.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Multilinear Regression | Decision Tree | Neural Networks |
| ME | 0.001 | -0.006 | 0.684 |
| RMSE | 0.23 | 0.29 | 0.724 |
| MAE | 0.181 | 0.225 | 0.687 |
| MPE | -0.318 | -0.5601 | 13.705 |
| MAPE | 4.321 | 5.3367 | 13.76 |
| Correlation | 0.92 | 0.87 | 0.91 |

Among the three models—Multilinear Regression, Decision Tree, and Neural Networks—Multilinear Regression demonstrates the most favorable performance based on various evaluation metrics. With low Mean Error (ME), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), this model indicates accurate predictions, while its high correlation of 0.92 suggests a strong linear relationship with the target variable. In contrast, the Decision Tree model shows reasonable predictive performance but with slightly higher percentage errors. Neural Networks, although capturing complex patterns with a high correlation of 0.91, exhibit significantly higher errors, indicating potential challenges in accuracy.

### Model Evaluation for Predicting the price of a used phone using final selected model.

**Model:**

**A close up of a text

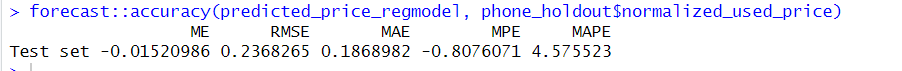
Description automatically generated**

**Evaluation:**

A screenshot of a computer code

Description automatically generated

The summary statistics for the predicted prices from the regression model on the holdout dataset reveal a distribution with a minimum of 2.742, a median of 4.386, and a maximum of 6.197. Correspondingly, the summary statistics for the actual normalized used prices in the holdout set range from a minimum of 2.268 to a maximum of 6.619, with a median and mean close to those of the predicted prices (4.407 and 4.361, respectively). The high correlation coefficient of 0.92 between the predicted and actual prices indicates a robust positive linear relationship, suggesting that the regression model effectively captures the variations in normalized used prices for phones in the holdout dataset. This alignment between summary statistics and the high correlation underscores the model's accuracy in predicting prices on the unseen holdout data.



The forecast accuracy assessment of the regression model on the holdout dataset reveals a generally reliable performance. The mean error (ME) is slightly negative (-0.0152), indicating a slight underestimation on average. The root mean squared error (RMSE) of 0.2368 suggests a moderate overall prediction error, showcasing the model's reasonable ability to capture variations in phone prices. The mean absolute error (MAE) is 0.1869, signifying a moderate level of accuracy in predicting absolute price values. Despite a minor tendency for underestimation reflected in the mean percentage error (MPE) of -0.8076%, the mean absolute percentage error (MAPE) at 4.5755% highlights the model's effectiveness in capturing relative price differences. These metrics collectively suggest that the regression model maintains a reasonably accurate forecasting performance on the unseen holdout data.

## Models Performance Comparison for Classification of device

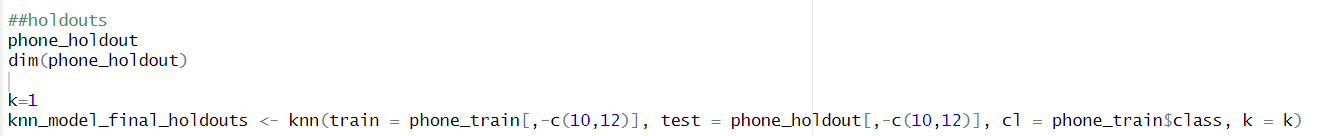
Here we performed 4 different models for classification of phone into high configuration, medium and low configuration phone which are multinomial regression, classification tree, naïve bayes and KNN. All the models has their own distinct characteristics. The confusion matrix analysis unveils specific insights into their performance. The multinomial regression model excels in sensitivity, capturing the nuances of both High and Low configurations. The classification tree demonstrates its interpretability by revealing decision rules, while Naïve Bayes leverages probabilistic assumptions. KNN, relying on proximity-based classification, contributes a different paradigm. The balanced accuracy metric gauges the trade-offs each model makes between precision and recall across classes. This holistic evaluation aids in selecting the most suitable model based on the specific priorities and requirements of the classification task.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Multinomial Logistic | Classification Tree | Naïve Bayes | KNN |
| Accuracy | 33% | 98% | 66% | 79% |
| Sensitivity | High: 91.87%  Medium: 1.25%  Low: 1.9% | High: 97.97%  Medium: 97.56%  Low: 99.17% | High: 80.89%  Medium: 18.53%  Low: 93.33% | High: 89.02%  Medium: 64.88%  Low: 81.25% |
| Specificity | High: 97.3%  Medium: 53.65%  Low: 51.23% | High: 100%  Medium: 98.56%  Low: 98.89% | High: 91.01%  Medium: 90.94%  Low: 67.73% | High: 95.51%  Medium: 86.83%  Low: 86.7% |
| Error Rate | 67% | 2% | 30% | 21% |

Based on all the metrics mentioned, classification tree is outperforming all other models. However, would choose KNN over classification tree assuming there might be a overfitting. KNN model is also performing well in terms of accuracy, sensitivity and specificity.

### Model Evaluation for Classification of phone using final selected model.

**Model:**



**Evaluation:**

A screenshot of a computer

Description automatically generated

The final selected model KNN, performs well in holdouts as well with the confusion matrix reveals a classification model's performance on predicting High, Medium, and Low configurations. With an overall accuracy of 80.75%, the model shows proficiency in correctly classifying instances. Class-specific metrics, including sensitivities (High: 88.62%, Medium: 69.76%, Low: 82.08%), and balanced accuracy (High: 96.18%, Medium: 86.01%, Low: 89.36%), provide a detailed understanding of the model's effectiveness in distinguishing between the different phone configurations. The Kappa statistic of 0.7106 indicates substantial agreement beyond chance, reinforcing the model's reliability.

The final selected models for predicting the price and to classify the device into high-medium-low configured device are multi linear regression, KNN which performed well in holdouts as well with low RMSE score, error measures and good accuracy.

# **CONCLUSION**

In conclusion, the integration of predictive pricing and classification models presents a significant boon for Business 'X' operating within the dynamic mobile phone market. By leveraging these models, the business gains a strategic advantage, enhancing its ability to navigate market intricacies effectively.

The predictive pricing model, particularly the multilinear regression approach, performs reasonably well which exhibits a high degree of accuracy, boasting a commendable 0.92 correlation with actual prices. This precision empowers Business 'X' to optimize inventory management, striking a balance between supply and demand to prevent overstocking or understocking. Additionally, it facilitates the establishment of competitive pricing strategies, ensuring profitability while attracting a broader customer base.

Furthermore, the classification model, led by the KNN algorithm with an impressive 80% accuracy rate, facilitates the categorization of used phones into high-medium-low configurations. This classification system enables tailored marketing strategies and personalized offerings, catering to the diverse needs and preferences of customers. By segmenting the market effectively, Business 'X' can refine its sales approaches, attracting customers with targeted promotions and enhancing overall customer satisfaction.

Ultimately, these predictive pricing and classification models serve as invaluable tools, empowering Business 'X' to make informed decisions, optimize operations, and stay abreast of market trends. By aligning with evolving consumer demands and market dynamics, the business secures a competitive edge in the thriving landscape of the used phone industry. Through strategic utilization of these models, Business 'X' stands poised for sustained success and growth in the ever-evolving mobile phone market.

Shruthi Bashetti

17th January 2024,

# **EXECUTIVE SUMMARY**

The mobile phone market is very dynamic in today’s world, for buyers and sellers it is very important to understand the factors influencing used smartphone pricing is crucial. Business ‘X’ is into the sales of a used phones, which needs a model that predicts the price of these phone. Accurate price predictions enable businesses to optimize their inventory by stocking phones at prices aligned with market demand. This helps prevent overstocking or understocking, ensuring a balanced and profitable inventory and deciding the profit margins. Price predictions inform marketing strategies, helping businesses design promotions, discounts, and advertising campaigns that resonate with the target market. This enhances the effectiveness of marketing efforts and attracts a larger customer base. Business also needs a model to classify these phones into high-medium-low configured phones. This classification system enables personalized offerings to customers based on their specific needs and preferences. Classification allows businesses to tailor marketing strategies and sales approaches based on the different phone configurations. This targeted approach helps attract customers with specific preferences and needs, increasing the chances of successful transactions.

In order to help the business with the needs, we built different models for predicting the price of a used phone using an analytical approach. The models can predict prices of used phones accurately. Out of which multilinear regression is outperforming other models built which gives 0.92 correlation with actual prices. To classify the phones into high-medium-low configuration phones, we built several models out of which KNN is outperforming which is giving almost 80% accuracy. So, the models built for these requirements will significantly benefit Business 'X' in navigating the intricacies of the dynamic mobile phone market. In conclusion, the integration of predictive pricing and classification models not only streamlines operations for Business 'X' but also empowers the business to make informed decisions, optimize inventory, set competitive prices, and enhance customer satisfaction. These models serve as valuable tools, aligning the business with market trends and ensuring a competitive edge in the ever-evolving landscape of the used phone industry.