**Used Device Price Prediction & Classification**

Project 2

Sreeja Reddy Singidi

03-20-2024

# **PROJECT 2: Used Device Price Prediction & Classification**

**TABLE OF CONTENTS**

[PROJECT 2: Used Device Price Prediction & Classification 2](#_Toc165748628)

[**LIST OF FIGURES** 4](#_Toc165748629)

[**LIST OF TABLES** 4](#_Toc165748630)

[1. INTRODUCTION 5](#_Toc165748631)

[2. PROJECT OVERVIEW 5](#_Toc165748632)

[3. BUSINESS UNDERSTANDING 6](#_Toc165748633)

[4. BUSINESS GOAL 6](#_Toc165748634)

[5. ANALYTICAL APPROACH 6](#_Toc165748635)

[6. ATTRIBUTE DEFINITION 7](#_Toc165748636)

[7. DATA UNDERSTANDING 10](#_Toc165748637)

[**7.1** **SUMMARY STATISTICS** 10](#_Toc165748638)

[**7.2** **CHECKING FOR MISSING VALUES** 10](#_Toc165748639)

[**7.3** **CHECKING FOR 0 VALUES** 11](#_Toc165748640)

[**7.4** **ATTRIBUTE ANALYSIS** 12](#_Toc165748641)

[**7.4.1** **DISTRIBUTION OF TARGET VARIABLE** 12](#_Toc165748642)

[**7.4.2** **CATEGORICAL ATTRIBUTES** 12](#_Toc165748643)

[**7.4.3** **NUMERICAL ATTRIBUTES** 14](#_Toc165748644)

[8. PREDICTOR ANALYSIS AND RELEVANCY 19](#_Toc165748645)

[8.1 ANALYSIS 19](#_Toc165748646)

[**8.1.1** **AVERAGE PRICE BY RAM** 19](#_Toc165748647)

[**8.1.2** **RELATION BETWEEN THE USED PRICE AND A NEW PRICE** 19](#_Toc165748648)

[**8.1.3** **CORRELATION BETWEEN THE NUMERIC DATA** 20](#_Toc165748649)

[**8.1.4** **PRICE PER BRAND** 21](#_Toc165748650)

[**8.1.5** **PRICE PER OS** 21](#_Toc165748651)

[8.2 RELEVANCY 22](#_Toc165748652)

[**8.2.1** **LASSO MODEL** 22](#_Toc165748653)

[**8.2.2** **BORUTA METHOD** 22](#_Toc165748654)

[**8.2.3** **STEPWISE REGRESSION** 23](#_Toc165748655)

[9 DATA TRANSFORMATION 24](#_Toc165748656)

[**9.1** **CHANGING COLUMN NAME** 24](#_Toc165748657)

[10 DIMENSION REDUCTION 24](#_Toc165748658)

[**10.1** **COMBINING NETWORK TYPE** 24](#_Toc165748659)

[11 DATA PARTITIONING METHODS 24](#_Toc165748660)

[NECESSITY 24](#_Toc165748661)

[DIFFERENT APPROACHES 24](#_Toc165748662)

[APPROACH IN THIS PROJECT 25](#_Toc165748663)

[IMPORTANCE 25](#_Toc165748664)

[12 TABLET AND PHONE DATASETS 25](#_Toc165748665)

[13 GOAL 1: PREDICTING THE PRICE OF A USED DEVICE 25](#_Toc165748666)

[**13.1** **DATA PREPARATION FOR REGRESSION MODELS** 26](#_Toc165748667)

[**Step 1:** 26](#_Toc165748668)

[**Step 2: Dividing the dataset into Phone and Tablets** 27](#_Toc165748669)

[**Step 3: Partitioning the datasets.** 28](#_Toc165748670)

[**13.2** **REGRESSION MODELS** 28](#_Toc165748671)

[**13.2.1** **LINEAR REGRESSION MODEL** 28](#_Toc165748672)

[**13.2.2** **STEPWISE REGRESSION** 28](#_Toc165748673)

[**13.2.3** **REGRESSION TREE** 28](#_Toc165748674)

[**Regression Models for Tablet Dataset** 29](#_Toc165748675)

[**13.3** **REGRESSION MODEL SELECTION** 30](#_Toc165748676)

[14 GOAL 2: CLASSIFYING USED DEVICE PRICE CATEGORIES 30](#_Toc165748677)

[**14.1** **DATA PREPARATION FOR CLASSIFICATION MODELS** 31](#_Toc165748678)

[**Step 1:** 31](#_Toc165748679)

[**Step 2: Removing Unnecessary Columns** 31](#_Toc165748680)

[**Step 3: Tablet and Phone Division** 31](#_Toc165748681)

[**Step 4: Partitioning the datasets.** 31](#_Toc165748682)

[**14.2** **CLASSIFICATION MODELS** 32](#_Toc165748683)

[**14.2.1** **LOGISTIC REGRESSION** 32](#_Toc165748684)

[**14.2.2** **DECISION TREE** 32](#_Toc165748685)

[**14.2.3** **K-NEAREST NEIGHBORS** 33](#_Toc165748686)

[15 NEW AND USED PRICE COMPARISON FOR PHONE AND TABLET DATASETS 33](#_Toc165748687)

[16 CONCLUSION 34](#_Toc165748688)

[17 BUSINESS RECOMMENDATIONS / FUTURE WORK 34](#_Toc165748689)

[EXECUTIVE SUMMARY 35](#_Toc165748690)

## **LIST OF FIGURES**

[Figure 1: Overview. 5](#_Toc165748691)

[Figure 2: Missing values. 10](#_Toc165748692)

[Figure 3: Distribution of target variable. 12](#_Toc165748693)

[Figure 4: Distribution of OS 12](#_Toc165748694)

[Figure 5: Distribution of Brands 13](#_Toc165748695)

[Figure 6: Distribution of networks. 13](#_Toc165748696)

[Figure 7: Distribution of screen size 14](#_Toc165748697)

[Figure 8: Distribution of rear camera. 14](#_Toc165748698)

[Figure 9: Distribution of front camera. 15](#_Toc165748699)

[Figure 10: Distribution of internal memory. 15](#_Toc165748700)

[Figure 11: Distribution of RAM 16](#_Toc165748701)

[Figure 12: Distribution of battery. 16](#_Toc165748702)

[Figure 13: Distribution of weight. 17](#_Toc165748703)

[Figure 14: Distribution of release year. 17](#_Toc165748704)

[Figure 15: Distribution of days used. 18](#_Toc165748705)

[Figure 16: Boxplot for numeric attributes 18](#_Toc165748706)

[Figure 17: Average Price by RAM 19](#_Toc165748707)

[Figure 18: Relation between used price and new price 20](#_Toc165748708)

[Figure 19: Correlation 20](#_Toc165748709)

[Figure 20: Price per brand 21](#_Toc165748710)

[Figure 21: Price per OS 21](#_Toc165748711)

[Figure 22: Lasso models 22](#_Toc165748712)

[Figure 23: Boruta Method 23](#_Toc165748713)

[Figure 24: Tablet- Regression trees 29](#_Toc165748714)

[Figure 25: Phone- Regression trees 29](#_Toc165748715)

[Figure 26: Phone- Classification Tree 32](#_Toc165748716)

[Figure 27: Price comparison. 34](#_Toc165748717)

## **LIST OF TABLES**

[Table 1: Attributes 9](#_Toc165723650)

[Table 2: Transforming OS 26](#_Toc165723651)

[Table 3: Transforming Network Type 26](#_Toc165723652)

[Table 4: Transforming Brand 27](#_Toc165723653)

[Table 5: Tablet and Phone Partitioning 28](#_Toc165723654)

[Table 6: Tablet- Regression Models 29](#_Toc165723655)

[Table 7: Phone- Regression Models 30](#_Toc165723656)

[Table 8: Phone Tablet Classification Partitioning 31](#_Toc165723657)

[Table 9: Tablet-train and test records for target 31](#_Toc165723658)

[Table 10: Phone- train and test records for target 31](#_Toc165723659)

[Table 11: Tablet- Classification Models 33](#_Toc165723660)

[Table 12: Phone- Classification Models 33](#_Toc165723661)

# **INTRODUCTION**

Used smartphone demand has increased significantly in today's dynamic world of technology and business, driven by several results like affordability, sustainability, and technological improvements, but it's still quite difficult to determine these device’s true worth among the many different features, and brands. "Predict & Win, is a gaming platform that aims to tackle this difficulty by providing consumers with an organized approach to understanding the complicated world of pricing for used smartphones. It also provides an engaging experience in estimating the value of used devices. Within this innovative environment, users explore the complex dynamics of the used mobile phone market in a creative environment, taking them on an interesting trip.

The platform will include an easy-to-use interface with simple controls for submitting predictions and monitoring results.

The gamification elements also encourage active involvement and skill improvement by rewarding users with points and badges which they can redeem when purchasing a used device.

# **PROJECT OVERVIEW**

A screenshot of a diagram

Description automatically generated

Figure 1: Overview.

# **BUSINESS UNDERSTANDING**

Predict and Win operates as an interactive gaming platform where users engage in estimating the resale price of used mobile phones. Additionally, the platform serves as a marketplace where users can purchase used devices, thereby offering a comprehensive solution for both gaming and e-commerce needs.

As an interactive gaming platform, it allows players to estimate the cost of reselling used mobile phones. To generate meaningful estimates, participants examine a variety of elements using an easy-to-use interface, including device characteristics, market trends, and historical data. These rough estimates provide the foundation for earning rewards that can be redeemed while purchasing a device, which increases user motivation and engagement.

Predictive analytics is applied in "Predict and Win" to create models that reliably forecast used mobile phone resale values. To guarantee that customers are presented with realistic estimation scenarios, these models are continuously improved through the application of sophisticated algorithms and machine learning approaches.

# **BUSINESS GOAL**

The primary business goal of "Predict and Win" revolves around revenue generation. Premium challenges can be purchased using participation fees as one source of revenue. To participate in high-stakes challenges with exclusive rewards, users may be required to pay an entry fee, which goes toward funding the platform. For customers who are prepared to pay a participation fee, this will provide premium challenges with greater stakes and exclusive rewards. These challenges make money through entrance fees and sponsorships, offering devices, or sponsored gifts from collaborating firms.

Additional revenue opportunities can be obtained through partnerships with industry partners, merchants, and device manufacturers. Through revenue-sharing agreements or sponsorship payments, these partners may support contests, feature new items, and raise brand recognition on the site.

# **ANALYTICAL APPROACH**

The analytical approach to build models that can be used to predict the used prices for the game includes the following:

Data analysis: The data should be first cleaned- handling the missing values and zero values. This part includes attribute analysis like the distribution of target variables, discovering patterns, finding outliers, and knowing the relationship between each variable.

Model building: We will be using classification and regression algorithms for used price categories and used price predictions respectively.

Model evaluation: In further steps, the best model will be picked- its accuracy and performance will be assessed based on that.

Deployment: After the testing is done, they can be deployed for used price recommendations and categories for used devices.

After constructing predictive models for estimating used smartphone prices and categories, these predictions become the backbone of the gaming experience. Customer’s predictions are based on these model-estimated predictions, and they receive prizes if their guess is within the expected range; if not, no additional points are awarded.

By providing rewards for precise prediction, this approach raises user involvement and promotes data-driven decision-making. It guarantees that the prediction models have a useful function within the game platform, providing users with an enjoyable and enlightening experience.

# **ATTRIBUTE DEFINITION**

1. device\_brand: This attribute shows the brand name of the smartphone like Apple, Infinix, Huawei, etc. This will help to know the manufacturer of the smartphone and also for any analysis that is brand specific. The type of the variable is character. ‘Others’ means brands that are not well-known or comparatively smaller brands.
2. os: This attribute tells which operating system is used in the phone like IOS, Android, Windows, or others which are not used much.
3. screen\_size: The attribute shows the screen size of the smartphones in inches which is a numeric type. It helps identify the physical dimensions of the display, which can improve the customer’s usability.
4. X4g: This is a binary attribute that shows whether the given smartphone supports a 4G network or not. If it supports 4g then yes, otherwise no.
5. X5g: This is similar to the above attribute which shows whether the given smartphone supports a 5G network or not. If it supports then yes, otherwise no.
6. rear\_camera\_mp: This is an important feature of the smartphone for youth as they focus on taking good pictures and videos. This shows how many megapixels the main camera which is on the back has. Megapixel just means how much detail the camera can capture. More the megapixel more the clarity the photos.
7. front\_camera\_mp: It is same as the back camera. This concentrates on the megapixels of the front-facing camera in smartphones which shows how well the front camera can capture selfies or self-portraits.
8. internal\_memory: It describes the amount of storage that can be used to store files, media, and applications that is numeric. It affects the cost, usability, and general performance of the phone. Customers frequently take internal memory size into account to know if a used smartphone is valuable and appropriate for their needs.
9. ram: Used smartphone must have Random Access Memory because it is essential for multitasking and temporary data storage. In this case - larger RAM capacities are preferable since they can perform multiple activities at once without experiencing any lag.
10. battery: The battery is measured in mAh (milliampere-hours) which tells how much power the phone can hold. For customers who use the phone all day, battery becomes essential. An increased battery life is linked with higher mAh values.
11. weight: Shows the weight of the phone in grams.
12. release\_year: Understanding the smartphone's release year affects its value in the market by allowing one to look at its technological innovations. To help with pricing and estimates of demand for used phones, the release year analyzes the trends and the understanding of customer preferences over time.
13. days\_used: This attribute shows the number of days the smartphone has been in use since purchase which may have an impact on its state. Provides insights into customer behavior related to the lifespan and replacement cycles of devices.
14. normalized\_used\_price: For determining the cost of used smartphones based on their characteristics and specifications is made by standardizing the values. Moreover, the models can successfully identify trends and patterns in the used smartphone market by using normalized prices, which helps buyers and sellers make smarter decisions.
15. normalized\_new\_price: These are the standardized values of the smartphone's original price based on the features. It adjusts the prices to the same range, to give fair and clear comparisons across different models and brands.

|  |  |  |
| --- | --- | --- |
| Attribute Name | Description | Type |
| device\_brand | Brand of the smartphone | Categorical |
| os | Operating system of the smartphone | Categorical |
| screen\_size | Size of the smartphone screen in cms | Numeric |
| X4g | Phone having 4g or no | Categorical (yes or no) |
| X5g | Phone having 5g or no | Categorical (yes or no) |
| rear\_camera\_mp | rear camera in megapixels | Numeric |
| front\_camera\_mp | front-facing camera in megapixels | Numeric |
| internal\_memory | Internal memory capacity of the smartphone in GB | Numeric |
| ram | Random Access Memory (RAM) of the smartphone in GB | Numeric |
| battery | The capacity of the smartphone's battery in milliampere-hours (mAh) | Numeric |
| weight | Weight of the smartphone in grams | Numeric |
| release\_year | Year of release of the smartphone | Numeric (Year) |
| days\_used | Number of days the smartphone has been used | Numeric |
| normalized\_used\_price | Normalized price of the used smartphone | Numeric |
| normalized\_new\_price | Normalized price of the original smartphone | Numeric |

Table 1: Attributes

# **DATA UNDERSTANDING**

The dataset consists of 3454 records with 15 variables that are the features of the used phone with the target variable.

## **SUMMARY STATISTICS**

Summary statistics will be relevant for numeric data because they show the mean, median, and quartiles of the variables with min-max values.

***A screenshot of a computer screen

Description automatically generated***

Code Snippet 1: Summary Statistics

## **CHECKING FOR MISSING VALUES**

There are 202 missing values in total and a huge number of missing values in the variable rear camera\_mp which is the back camera as seen in Figure 1.

A graph with blue squares and pink squares

Description automatically generated

Figure 2: Missing values.

Investigating the highest missing values variable i.e, rear camera:

Missing values in rear camera relation with year:

When going through the dataset we can notice that 122 records are missing from the *year 2022* out of 277 records in that particular year. This might be because they did not record these details consistently or brands might not have focused much on giving details about the latest phone’s rear cameras.

Relation with Operating System:

All 179 records are associated with Android phones.

**Replacing missing values by group medians:**

The dataset is initially grouped by brand, creating separate groups for each unique brand. Within each group, the median value for specific columns is calculated to handle missing values.

However, even after this process, 10 records remain missing. Upon investigation, it's discovered that all these missing records belong to the brand "Infinix." Infinix models from the years 2017 to 2020 are known to have rear cameras. This discrepancy suggests that the missing values for the rear camera are likely due to data entry errors or incomplete data for Infinix models. Given this, these 10 records are considered unreliable and can be safely removed from the dataset.

## **CHECKING FOR 0 VALUES**

Only the front camera has 39 zero values as seen in Code snippet 2.

Analyzing the reason for these 0 values - all 0 values are from the Nokia brand that supports ‘Other’ operating systems apart from Android and Windows with no 5g network. The reason could be that fewer common models did not prioritize the front camera whereas the manufacturers might have designed the mobile for other specific features rather than the front camera.

Nokia had its own Symbian OS but later shifted to Android and Windows due to external factors. So, this can be one of the reasons for these 0 values or the records might be missing. For now, those zero values need not be changed.

A close-up of a computer code

Description automatically generated

Code Snippet 2: Zero values.

## **ATTRIBUTE ANALYSIS**

### **DISTRIBUTION OF TARGET VARIABLE**

Blue to red color in Figure 2, shows the relative frequency of used device prices inside each bar, which reflects a certain range of normalized values. Bars with red color show areas with higher concentrations of used phones. The highest prices go between 4-5 from the plot below. The shape shows the distribution of prices.

A graph of a normalized used price

Description automatically generated

Figure 3: Distribution of target variable.

### **CATEGORICAL ATTRIBUTES**

#### **OPERATING SYSTEM**

According to Figure 3, Android has the highest count of used smartphones than others. This shows that this dataset sees higher sales in Android due to the availability of different used device prices and because of customer preferences.

A graph of operating system

Description automatically generated

Figure 4: Distribution of OS

#### **DISTRIBUTION OF BRANDS**

We can analyze the features and pricing of top brands. By doing this it is possible to compare the best brands to the others. We can also analyze predictive models that focus on top brands to check if they can improve the accuracy of models. The top brands from Figure 4, are Others (brands apart from this dataset), Samsung, Huawei, LG, and Lenovo.

A graph of different colored bars

Description automatically generated

Figure 5: Distribution of Brands

#### **NETWORKS TYPE**

From Figure 5 it indicates the majority of smartphones support 4g. We can say that 5G is more recent and has not yet become as popular in the used smartphone market. By identifying instances where devices lack 4G and 5G capabilities, we gain an understanding of the distribution of network technologies in the market. There are higher chances of buyers considering “other” networks like 3g, and 2g than 5g.

A graph of a network distribution

Description automatically generated

Figure 6: Distribution of networks.

### **NUMERICAL ATTRIBUTES**

#### **DISTRIBUTION OF SCREEN SIZE**

The distribution of screen sizes indicates that 12-15 inches is the range in which most mobile devices fall, indicating a widely accepted standard size in the market. With a screen size of thirty inches as seen in Figure 6, there is an interesting outlier that suggests a gadget that probably might be a tablet.

*A graph of green bars

Description automatically generated*

Figure 7: Distribution of screen size

#### **DISTRIBUTION OF REAR CAMERA**

The examination of back camera megapixels indicates a widespread tendency toward 13-megapixel cameras. But significant percentages also include 6 and 8-megapixel cameras, showing variation in camera specifications. In this dataset, no phone has megapixels between 25 to 40 possibly indicating a lack of devices with higher-resolution cameras with this dataset.

A graph of a camera size

Description automatically generated

Figure 8: Distribution of rear camera.

#### **DISTRIBUTION OF FRONT CAMERA**

Notably, 6–7-megapixel front cameras are the most popular option, indicating possible availability in this market. Furthermore, the existence of phones with front cameras over 30 megapixels suggests that there is a wide range of options available to suit customers' demands for different levels of performance and quality. These kinds of insights are crucial for buyers and sellers to successfully navigate the used phone market.

A graph of a camera size

Description automatically generated

Figure 9: Distribution of front camera.

#### **DISTRIBUTION OF INTERNAL MEMORY**

The majority of used smartphones in the dataset possess internal memory capacities below 100GB, indicating common storage configurations. However, the presence of a record with 1000GB of internal memory suggests an outlier observation.

A graph with green and black bars

Description automatically generated

Figure 10: Distribution of internal memory.

#### **DISTRIBUTION OF RAM**

According to an analysis of the RAM distribution, only a small percentage of smartphones have 8GB of RAM, while the vast majority of devices have 4 GB. This finding shows that consumers strongly prefer smartphones with 4GB of RAM.

A graph with green rectangular bars

Description automatically generated

Figure 11: Distribution of RAM

#### **DISTRIBUTION OF BATTERY**

Only a small percentage of smartphones have battery capacities larger than 6000 mAh. Most used smartphones have battery capacities between 2000 and 4500 mAh. This finding highlights the widespread practice of smartphones having batteries that fall within a specific capacity range, which is important information for buyers and sellers on the trade-in platform to know when evaluating aspects like battery life and performance.

A graph of battery

Description automatically generated

Figure 12: Distribution of battery.

#### **DISTRIBUTION OF WEIGHT**

It can be seen from Figure 12, that a large percentage of the smartphones in the dataset weigh between 140 and 190 grams and a phone has 850 grams as the highest value.

A graph of a weight

Description automatically generated

Figure 13: Distribution of weight.

#### **DISTRIBUTION OF RELEASE YEAR**

The years 2013 and 2014 show the highest frequency of smartphone purchases, indicating an increase in demand from consumers during this time. This trend may be explained by the notable developments in smartphone technology over the past several years.

The least number of purchases, however, was made in 2020, suggesting a possible drop in customer interest or market saturation. This decline in purchases could be because of the **COVID** pandemic.

A graph of a number of green bars

Description automatically generated with medium confidence

Figure 14: Distribution of release year.

#### **DISTRIBUTION OF DAYS USED**

Most of the phones in the dataset were in use for 600–900 days after purchase, which suggests that consumers often keep their phones for this amount of time. Understanding the distribution of days used can help in assessing the depreciation rate of smartphones over time.

A graph of green and black bars

Description automatically generated

Figure 15: Distribution of days used.

#### **BOXPLOT FOR NUMERIC ATTRIBUTES**

In the interquartile range (IQR), the median is indicated by the horizontal line inside the box, and the whiskers extend to the minimum and maximum values within 1.5 times the IQR from the first and third quartiles, respectively, for each box plot that shows the distribution of values for a particular attribute. This visualization provides insights into the distribution and variability of numeric attributes in the dataset.

A screenshot of a computer

Description automatically generated

Figure 16: Boxplot for numeric attributes

# **PREDICTOR ANALYSIS AND RELEVANCY**

# **ANALYSIS**

### **AVERAGE PRICE BY RAM**

The line graph (Figure 16) shows the relationship between RAM capacity and the average price of used devices. It shows a continuous increase up to a RAM capacity of 12 with slight decreases in the price when RAM was high. This graph is useful for understanding how RAM capacity influences the price of used devices. Taking the average ensures that the graph represents a generalized trend across different RAM capacities rather than individual data points, providing a clearer picture of the overall relationship.

A graph of different colored bars

Description automatically generated

Figure 17: Average Price by RAM

### **RELATION BETWEEN THE USED PRICE AND A NEW PRICE**

The scatter plot indicates that they have a positive correlation, which means that as the normalized new price increases, the normalized used price tends to increase. This is to be expected because phones with higher new prices will probably also cost more when used. Knowing how new and used prices relate to one another can help customers decide which one to buy. They have a better chance to decide whether to buy new or used phones depending on their financial restrictions since they can predict how a smartphone's resale value may fluctuate over time.

A graph showing a normalized price

Description automatically generated

Figure 18: Relation between used price and new price

### **CORRELATION BETWEEN THE NUMERIC DATA**

Strong correlations were found between predictors, such as battery weight and weight-screen size, suggesting possible relationships between these features. On the other hand, internal memory-screen size and rear camera-5G showed lower correlations, indicating a weaker relationship between these factors. By understanding these correlation patterns, model-building procedures can be improved.

A diagram of a device

Description automatically generated with medium confidence

Figure 19: Correlation

### **PRICE PER BRAND**

The dataset contains a roughly uniform representation of smartphones across various price ranges which means there isn't a significant skew towards higher or lower-priced devices, and smartphones are distributed relatively evenly across different price categories.

A graph of a company brand

Description automatically generated

Figure 20: Price per brand

### **PRICE PER OS**

The graph (Figure 20) suggests that iOS devices have a higher value in the market, likely due to factors such as brand reputation. On the other hand, Android devices, while still popular, tend to be priced lower than iOS devices, reflecting their broader availability across a range of price points and manufacturers. Other operating systems, which are less widely used by the public, have minimal representation in the market and consequently do not attract significant demand or make high prices.

A graph of a bar chart

Description automatically generated with medium confidence

Figure 21: Price per OS

# **RELEVANCY**

To build precise predictive models, it is essential to identify and rank the most significant predictors. This can be performed by *feature selection*.

### **LASSO MODEL**

LASSO will consider irrelevant predictor’s coefficients to exactly zero to show variable selection. By eliminating predictors from the model that have little to no impact on the used price, LASSO effectively modifies the data. The predictors whose coefficients are non-zero after regularization are regarded as significant and are kept in the model. This model automatically considers all categorical attributes into dummies.

It is concept- oriented which gives better results for feature selection by shrinking the errors. So, this is considered an important method.

The significant predictors using this method are:



Code Snippet 3: Lasso important predictors.

Using cv.glmnet: Using glmnet:

|  |  |
| --- | --- |
| **A graph with numbers and a red dotted line  Description automatically generated**  Figure 22: Lasso models | A graph of a number  Description automatically generated with medium confidence |

### **BORUTA METHOD**

A wrapper-based feature selection technique called Boruta analyzes each predictor's significance. To determine their significance iteratively balance the weights of shadow attributes (random noise) to real features. In the end, Boruta classifies features as "Confirmed," "Tentative," or "Rejected" according to how important they are.

Every predictor appears to be significant, according to the Boruta feature selection plot. We'll use the random forest to verify the significance of this finding.

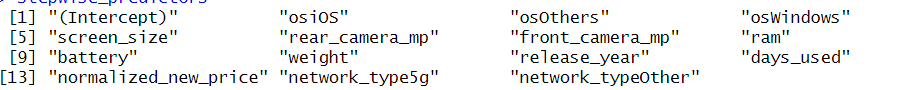
**A graph with green and blue squares

Description automatically generated**

Figure 23: Boruta Method

### **STEPWISE REGRESSION**

It is a statistical technique where predictors are gradually added or removed accordingly.

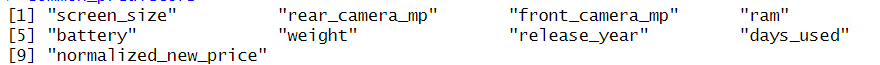


Code Snippet 4: Stepwise Regression predictors.

The common predictors found in all models after using these feature selection techniques are believed to be the most important variables for predicting the normalized used price of smartphones.

Intersecting the predictors with these models can help identify the important common features which can be seen in Code snippet 5.

These common features are likely to have a significant impact:



Code Snippet 5: Common significant predictors.

However, the final decision on which predictors to include in the models will be based on thorough evaluation during model building. Both Lasso and stepwise have similar important predictors. We'll compare models built with all predictors against those built with the identified important predictors using stepwise to make correct decisions about feature selection and model performance.

# **DATA TRANSFORMATION**

## **CHANGING COLUMN NAME**

The data transformation phase involves modifications to enhance readability and usability. The column name “device\_brand” can be changed to “brand”.

# **DIMENSION REDUCTION**

## **COMBINING NETWORK TYPE**

The 4G and 5G network separate columns are combined into a single column called "network\_type," which simplifies the dataset. It is a well-established concept that devices with 5G connections also have 4G connectivity by default. In addition to the differentiations between 4G and 5G networks, there is also a further category called "Other." This group includes situations in which a network does not support 4G or 5G and hence falls under the "Other" category.

# **DATA PARTITIONING METHODS**

# **NECESSITY**

Data partitioning is essential to predict models accurately. By dividing the dataset, we can train models on a portion of the data, assess their efficiency, and ensure the models operate well when applied to new, untested data. By preventing overfitting, optimizing model performance, and offering trustworthy smartphone pricing predictions, this method improves the effectiveness and reliability of the gaming platform.

# **DIFFERENT APPROACHES**

Data partitioning can be done in different methods:

**Train-test division**: The process of train-test partitioning involves splitting the dataset into two subsets: a training set for model training, and a testing set for model evaluation. This is expanded by train-valid-test partitioning, which includes a validation set for optimizing model parameters and hyperparameters for improved generalization and performance on new data which is divided.

**Cross Validation**: It is a resampling strategy that divides the dataset into several subsets of training and validation sets iteratively to assess the performance of the models. Training and testing a model on several subsets of the data lowers the likelihood of overfitting and offers a more reliable evaluation of the model's performance, which shows how well the model will generalize to new data.

# **APPROACH IN THIS PROJECT**

 The straightforward train-test split approach will be used in this project, allocating **70%** of the dataset for training and **30%** for testing. There is a significant amount of data to train the models, with 70% of the data. This gives the models enough information to identify patterns and connections in the data. Considering 30% of the data for testing ensures that we have a sufficient dataset to evaluate the model's performance accurately. This helps assess how well the trained models generalize to new data that is test data.

# **IMPORTANCE**

The main purpose is to create prediction algorithms that can precisely calculate used device prices for new, untested data. When a model learns irrelevant patterns or noise from training data, it becomes overfitted and performs negatively on the data. A 70-30 split reduces the chance of overfitting and increases the models' capacity for generalization by providing an adequate quantity of data for training. Overall, this partitioning facilitates the development of reliable predictive models that can be integrated into the "Predict & Win" gaming platform, offering users accurate estimations of used device prices.

# **TABLET AND PHONE DATASETS**

Screen sizes in the dataset range from 5 cm to 30 cm, with bigger sizes possibly referring to tablets instead of smartphones. It is essential to distinguish between these two device groups to guarantee precision and clarity in our prediction models. This classification will enable more accurate predictive modeling based on the unique features of each type of device.

So, the dataset division will be seen in further steps.

# **GOAL 1: PREDICTING THE PRICE OF A USED DEVICE**

The platform enables customers to estimate the resale worth of the devices with accuracy by projecting the price of a used mobile. This provides insightful information about the potential value of their devices, which improves the user experience overall. Since "Predict & Win" revolves around estimating device costs, precise forecasts add to the gamification element of the program. As part of the game, players guess prices and aim to make the most accurate predictions and win rewards.

Regression models are used for this purpose and to build the regression models, it is important to start by taking the entire data into a new dataset and making necessary transformations before building the models so that the original data is not disturbed and is intact. This involves the following steps:

## **DATA PREPARATION FOR REGRESSION MODELS**

To build the regression models, it is important to start by taking the entire data into a new dataset and making necessary transformations before building the models so that the original data is not disturbed and is intact. This involves the following steps:

**Step 1:** **Factoring variables**

Transforming categorical variables into numerical representations, we allow simpler interpretation and analysis within our regression models, enhancing the effectiveness of the Predict & Win gaming platform.

**Operating System (OS):**

|  |  |
| --- | --- |
| Original | Transformed |
| Android | 1 |
| iOS | 2 |
| Windows | 3 |
| Others | 4 |

Table 2: Transforming OS

**Network Type:**

|  |  |
| --- | --- |
| Original | Transformed |
| 4g | 1 |
| 5g | 2 |
| Other | 3 |

Table 3: Transforming Network Type

**Brand:**

|  |  |
| --- | --- |
| Original | Transformed |
| Honor | 1 |
| Others | 2 |
| HTC | 3 |
| Huawei | 4 |
| Lava | 5 |
| Lenovo | 6 |
| LG | 7 |
| Meizu | 8 |
| Micromax | 9 |
| Motorola | 10 |
| Nokia | 11 |
| OnePlus | 12 |
| Oppo | 13 |
| Realme | 14 |
| Samsung | 15 |
| Vivo | 16 |
| Xiaomi | 17 |
| ZTE | 18 |
| Apple | 19 |
| Asus | 20 |
| Coolpad | 21 |
| Acer | 22 |
| Alcatel | 23 |
| Blackberry | 24 |
| Celkon | 25 |
| Gionee | 26 |
| Google | 27 |
| Karbonn | 28 |
| Microsoft | 29 |
| Panasonic | 30 |
| Sony | 31 |
| Spice | 32 |
| Xolo | 33 |

Table 4: Transforming Brand

### **Step 2: Dividing the dataset into Phone and Tablets**

Dividing the dataset based on screen sizes into two categories, namely phones and tablets, is a crucial step in enhancing the predictive capabilities of Predict & Win. An average phone size might be up to 18 cm and beyond that can be considered as a tablet devices.

### **Step 3: Partitioning the datasets.**

|  |  |  |
| --- | --- | --- |
|  | Phone | Tablet |
| Train | 70%- 2225 records | 70%- 189 records |
| Test | 30%- 950 records | 30%- 80 records |

Table 5: Tablet and Phone Partitioning

## **REGRESSION MODELS**

Regression models are essential to the "Predict and Win" platform because they help users make decisions, drive gameplay factors, and increase user engagement in addition to giving consumers accurate pricing forecasts.

### **LINEAR REGRESSION MODEL**

Understanding the linear relationship between predictors and the target variable can be done with the help of linear regression. It offers understandable coefficients that show the intensity and direction of each predictor's association with the target. The model is built on training data and its performance is evaluated on test data.

### **STEPWISE REGRESSION**

To create regression models that are easier to understand, and more straightforward, stepwise regression helps in the selection of the most relevant predictors. Here the backward method is used. Choosing only those predictors that have a meaningful impact on predicting the target variable, avoids overfitting and enhances model generalization.

After the linear regression model is built, it is useful to evaluate the model's performance and consider the significant variables that can be compared with the other multiple linear models.

### **REGRESSION TREE**

Regression trees provide insights into the decision-making process and are also simple to understand and visualize. Pruning the regression tree using the best complexity parameter (cp) will give better results. Figure 23 and Figure 24 show the tree plots.

**Regression Tree - Tablet Dataset:**

Pruned Tree with all predictors: Tree with significant predictors:

|  |  |
| --- | --- |
| A diagram of a graph  Description automatically generated  Figure 24: Tablet- Regression trees | A diagram of a graph  Description automatically generated |

**Regression Tree - Phone Dataset:**

Pruned Tree with all predictors: Tree with significant predictors:

|  |  |
| --- | --- |
| A diagram of a computer generated image  Description automatically generated with medium confidence  Figure 25: Phone- Regression trees | A diagram of a graph  Description automatically generated with medium confidence |

## **Regression Models for Tablet Dataset**

|  |  |
| --- | --- |
| Models | Evaluation MAE |
| Linear Regression | 0.1879 |
| Linear Regression with significant predictors | 0.1810 |
| Stepwise Regression | 0.1860 |
| Regression Tree | 0.2149 |
| Regression Tree with significant predictors | 0.2119 |

Table 6: Tablet- Regression Models

#### **Regression Models for Phone Dataset**

|  |  |
| --- | --- |
| Models | Evaluation MAE |
| Linear Regression | 0.1781 |
| Linear Regression with significant predictors | 0.1777 |
| Stepwise Regression | 0.1777 |
| Regression Tree | 0.2278 |
| Regression Tree with significant predictors | 0.2232 |

Table 7: Phone- Regression Models

## **REGRESSION MODEL SELECTION**

**Tablet Dataset:** Among the models evaluated, **Linear Regression with Significant Predictors** exhibits the lowest MAE (0.1810) as seen in Table 6. This indicates that this model provides the most accurate predictions for used tablet device prices.

**Phone Dataset:** Both Linear Regression with Significant Predictors and Stepwise Regression show identical MAE values (0.1777) as seen in Table 7. Given this, selecting **Linear Regression with Significant Predictors.**

By selecting the most suitable regression model for each dataset, Predict & Win ensures more precise pricing estimations for both used tablets and phones. Efficient model selection also allows to allocation of resources effectively, focusing on the development and deployment of the most effective regression models. This optimization maximizes the platform's predictive capabilities while minimizing operational costs.

**Classifying the predicted price:**

After selecting the best regression models for phones and tablets, we can create a new column to show the predicted price categories high or low for easy understanding. Adding a new column with predicted price categories enhances the user experience within the platform. Users can quickly grasp the estimated value of their devices and make informed decisions regarding their participation in challenges or purchases on the platform.

We can also build classification models to predict the used price categories. As part of the **project's requirements to build classification models** for predicting the price categories of used devices, the following steps can be undertaken to achieve the classification goal.

# **GOAL 2: CLASSIFYING USED DEVICE PRICE CATEGORIES**

Categorizing predicted prices provides valuable market insights for both users and platform administrators. By analyzing the distribution of predicted price categories over time, administrators can identify trends, patterns, and changes in the used device market. This information can be used to improve pricing algorithms, update prediction models, and enhance the overall effectiveness of the platform. Classification also allows users to earn rewards or points based on the accuracy of their predicted price range category.

## **DATA PREPARATION FOR CLASSIFICATION MODELS**

To build the classification models, it is important to start by taking the entire data into a new dataset and making necessary transformations before building the models so that the original data is not disturbed and is intact. This involves the following steps:

**Step 1: Create a new column for the price category.**

Categorize the used price into two categories: "low" and "high". The threshold for categorization is based on the normalized used price, with values below 4.0 categorized as "low" and the rest as "high". This step is crucial as it defines the target variable for the classification models.

**Step 2: Removing Unnecessary Columns**

Variables like normalized used and new prices are not needed for classification so can be removed from the dataset. This step ensures that only relevant features are retained for model training, reducing noise, and improving model performance.

**Step 3: Tablet and Phone Division**

The dataset is divided into two subsets based on the screen size of the devices: tablets and phones. This division allows for separate modeling of tablet and phone devices, considering their characteristics and user preferences.

### **Step 4: Partitioning the datasets.**

|  |  |  |
| --- | --- | --- |
|  | Phone | Tablet |
| Train | 70%- 2224 records | 70%- 189 records |
| Test | 30%- 951 records | 30%- 80 records |

Table 8: Phone Tablet Classification Partitioning

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Used price category in Tablet | Train | | Test | |
| High | Low | High | Low |
| 186 | 3 | 79 | 1 |

Table 9: Tablet-train and test records for target

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Used price category in Phone | Train | | Test | |
| High | Low | High | Low |
| 1666 | 558 | 713 | 238 |

Table 10: Phone- train and test records for target

From Table 9 and Table 10, it's evident that there's a significant class imbalance between the "High" and "Low" categories in both the training and test datasets. This class imbalance can pose challenges for classification models because they may become biased towards predicting the majority class which is “High” due to its higher frequency. As a result, the model's performance metrics may not accurately reflect its ability to generalize to new data, especially for the minority class.

Despite the observed imbalance in the distribution of the used price categories, **we will proceed with building** the classification models using these partitions.

## **CLASSIFICATION MODELS**

### **LOGISTIC REGRESSION**

The glm() model is well suited for classification problems that give probabilities of belonging to a particular class. This is critical in a marketing environment where knowing the likelihood of a response is useful. The implementation and interpretation of logistic regression are made easier by its computing efficiency and simplicity.

### **DECISION TREE**

Decision trees can be used to identify the most significant features (e.g., brand, screen size, RAM) that contribute to the classification of devices into low or high-price categories. By analyzing these decision paths, users can gain insights into the factors affecting device pricing, enhancing their understanding of the market dynamics. Figure 25 shows the pruned tree for Phone Dataset.

A diagram of a cell phone

Description automatically generated

Figure 26: Phone- Classification Tree

### **K-NEAREST NEIGHBORS**

KNN can be utilized to classify devices into price categories based on their similarity to other devices in the dataset. By considering the characteristics of neighboring devices, KNN provides a flexible and adaptable approach to price category prediction, that shows varying market trends and user preferences.

#### **Classification Models for Tablet Dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| Models | Accuracy | Sensitivity | Specificity |
| Logistic Regression | 5% | 3.8% | 1 |
| KNN | 98.75% | 100% | 0 |

Table 11: Tablet- Classification Models

#### **Classification Models for Phone Dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| Models | Accuracy | Sensitivity | Specificity |
| Logistic Regression | 10% | 6% | 23% |
| Classification Tree | 89.17% | 95.93% | 68.91% |
| KNN | 88.01% | 93.41% | 71.85% |

Table 12: Phone- Classification Models

Tables 11 and 12 show the performance of the models for both tablet and phone datasets.

**These models are included as part of the project requirements but will not be utilized. Instead, the focus will be on categorizing the predicted used prices for better understanding and user engagement.**

# **NEW AND USED PRICE COMPARISON FOR PHONE AND TABLET DATASETS**

The gaming platform acquires useful insights into the distribution of pricing trends by calculating the percentage of new and used prices within each dataset. Users can discover a great deal about the current pricing dynamics in the tablet and phone marketplaces from these percentages.

The percentage of used devices in both datasets as shown in Figure 26, suggests that customers may be price-sensitive and might be willing to purchase used devices to save money. Therefore, offering affordable gaming options and promotions for used devices could attract more customers to the platform.

**A comparison of a pie chart

Description automatically generated**

Figure 27: Price comparison.

# **CONCLUSION**

In conclusion, the "Predict & Win" project has successfully developed predictive models for estimating used device prices, enhancing user engagement and market insights within the gaming platform. By utilizing these models, the platform increases user satisfaction and engagement by giving consumers useful market data in addition to precise pricing estimates. The platform will continue to be a reliable and entertaining place to anticipate and win prizes based on used device values if models and user engagement tactics are continuously improved.

# **BUSINESS RECOMMENDATIONS / FUTURE WORK**

The platform can focus on several avenues to further enhance the platform's predictive capabilities and user engagement. This may include expanding the dataset to include a wider range of device attributes and market factors for more comprehensive analysis and implementing user feedback mechanisms to continuously improve model accuracy and relevance to user needs. Additionally, exploring innovative gamification features and partnerships with device manufacturers or retailers could further enrich the user experience and drive platform growth.

# **EXECUTIVE SUMMARY**

**NAME**: Sreeja Reddy Singidi

**DATE**: 03-20-2024

**OPPORTUNITIES:**

The "Predict & Win" project offers an exciting chance to transform the market for used devices. Through the integration of gamification, e-commerce, and predictive analytics, the project aims to provide consumers with an interesting and unique platform for assessing the worth of their devices. Moreover, the project's focus on predictive analytics provides an opportunity to deliver valuable insights into market trends and pricing dynamics. Overall, the "Predict & Win" project represents a promising opportunity to create a dynamic and innovative platform that adds significant value to the used device market.

**SOLUTIONS:**

To achieve accurate price estimations for used smartphones and tablets on the Predict & Win platform, an approach is proposed. This involves implementing predictive analysis techniques such as regression modeling and classification algorithms to predict device prices based on various features. Additionally, by focusing on user engagement strategies and continuous improvement, the platform can establish itself as a leader in the device prediction space, capturing a significant share of the market and driving revenue growth.