Predicting Smartphone Prices with Regression Models

FA24 | Section 001 | Koehler

# Introduction

Smartphones have become a near ubiquitous good in today's world. Although markets are often dominated by a few brands, smartphones exist in many forms, sizes and, of course, prices. This report seeks to predict the prices of smartphones based on their different attributes, both physical (i.e., RAM, storage, color and whether it is attached to a cell phone contract) and non-physical (i.e., model and brand). The 'model' of the smartphone is considered a non-physical attribute because it is a label associated with the phone's design and 'newness', rather than a tangible feature.

This predictive model will be valuable for consumers to make informed financial decisions about what attributes they should look for in a phone to minimize the price they pay, including whether some of those attributes are more important than others in determining price. By combining predicted smartphone prices with the features that are most personally important to them, consumers can thus better understand which type of smartphone is most suitable for them.

Additionally, smartphone companies may find this model useful in comparing their pricing strategy with those of their competitors. This can inform adjustments to smartphone pricing according to evolving market benchmarks. This model is thus valuable to both consumers, firms and researchers alike.

## Dataset Description

This dataset was [sourced from Kaggle](https://www.kaggle.com/datasets/juanmerinobermejo/smartphones-price-dataset/data) and contains data of different smartphones sold on the Spanish platform, PC Componentes. It has 1816 entries, each representing a different smartphone listed for sale on PC Componentes, and 8 columns, defined as below. Attributes in these columns will be used to predict the target variable, Final Price. This dataset is limited in how there is no date given for when this dataset was created. As a result, it is not possible to place each smartphone entry within the context of how 'new' its model is, which is often a significant factor in determining price. The Kaggle post was updated in April 2024, however, so this project will assume that this is when the data was scraped.

|  |  |  |
| --- | --- | --- |
| Variable | Class | Description |
| Smartphone Name | object | The unique identifier or model name of the smartphone |
| Brand | object | Smartphone brand |
| Model | object | Smartphone brand model |
| RAM (Random Access Memory) | float64 | The amount of memory available for multitasking and phone performance (GB) |
| Storage | float64 | Capacity of the smartphone (GB) |
| Color | object | Color of the smartphone |
| Free | object | Whether or not the smartphone is attached to a cellphone contract |
| Final Price | float64 | The cost of the smartphone ($) |

# Preliminary Data Exploration

## Descriptive Statistics

A screenshot of a cell phone

Description automatically generated

Figure 1: Descriptive statistics for smartphone prices.

A comparison of a graph

Description automatically generated

Figure 2 (left) and 3 (right): Boxplot and histogram of smartphone prices.

Phones are priced within a range of $60.46 to $2,271.28, with an overall average price of $492.17. As in Figure 3, smartphone prices are positively skewed, meaning that most phones are priced lower around the median of $349.99. Nevertheless, there are several highly priced smartphones that exist as outliers outside the interquartile range, as seen in Figure 2.

A screenshot of a cell phone

Description automatically generated

Figure 4: The 10 most expensive smartphones in this dataset.

These high outliers were further explored by grouping smartphones by brand. It was found that Apple is the most expensive brand on average, with SPC and Qubo being the cheapest according to their mean prices. The 10 most expensive phones belong to Honor, Samsung and overwhelmingly, Apple (Figure 4). These smartphones represent the high outliers pictured in Figure 2.

## Free Cellphone Plan?

A blue circle with a number of percentages

Description automatically generated

Figure 5: Proportion of smartphones with a free cellphone plan to those without.

Figure 5 shows how most smartphones in the dataset included a free cellphone plan, with only 2.4% being offered without one. When graphed against each other in Figure 6, phones without a free cellphone plan formed two main clusters: one low and near the median, and the other around the upper whisker of the boxplot in Figure 2.

A graph of a number of blue and green bars

Description automatically generated

Figure 6: Histogram of prices split by if they have a free cellphone plan.

## Color

A graph of different colored bars

Description automatically generated

Figure 7: Histogram of prices split by color.

Figure 7 has largely the same shape as the overall histogram of smartphone prices in Figure 3, with no one color significantly affecting the distribution of the data. This is likely because colors are valued subjectively by consumers and may not impact pricing decisions, since these subjective preferences are difficult to analyze. In an early multiple linear regression performed with all variables, Color had a negative feature importance (Appendix A), which indicated that it was irrelevant or harmful to the model’s performance.

## Storage and RAM

A diagram of a storage data

Description automatically generated with medium confidence

Figure 8: Scatterplot of smartphone storage vs price.

A graph with blue dots

Description automatically generated

Figure 9: Scatterplot of smartphone RAM vs price.

As in Figure 6, smartphones have storage in increments that range from 2GB to 1TB. Both scatterplots for storage and RAM show a general positive trend, where the higher the storage or RAM, the higher the smartphone price. Nevertheless, this relationship is only moderate, as storage and RAM increments range widely in price. Phones with 1TB of storage are almost always higher priced. Phones can also be priced low at all storage and RAM increments, suggesting that having a smartphone with high storage/RAM specifications likely does not give brands a competitive advantage.

# Modeling and Interpretations

As all features except Color were seen to have some relationship with smartphone prices in this dataset, they were used in the model to predict prices. To get the most accurate and appropriate predictions, I built three different models and compared these against a baseline. I aim to choose the model that performs best with the lowest mean squared error (MSE), as this is a regression task with numerical and continuous target data. For each model, I used a train-test-split to train my model on 80% of the data, testing it on the remaining 20%.

## Baseline Model

A baseline model was used to assess the performance of other models. It uses the mean price, $492.18, as a blanket prediction for all smartphones. **The baseline model had an MSE of 158,799.40**.

## Multiple Linear Regression Model

A multiple linear regression (MLR) was chosen as a possible model because smartphone prices are a continuous variable. From initial data exploration, it also seems that multiple factors have a relationship with the target variable, and their combined influence can be modelled using an MLR. Early visualizations of the relationships between features (such as storage and RAM) and price also displayed a linear relationship.

A screenshot of a computer program

Description automatically generated

Figure 10: Pipeline for the multiple linear regression model.

A pipeline was created by first preparing the data for regression. This involved encoding categorical variables using OneHotEncoder and using a StandardScaler on all other features. The encoded training data was then fed through a linear regression when a fit was called.

Both the training and test data performed better than baseline, with an **MSE of 9,954.23 and 46,855.46 respectively**. This was expected since the MLR was able to utilize the features to make more subtle predictions, rather than just using the mean smartphone price as the baseline did.

A white table with black text and numbers

Description automatically generated

Figure 11: Importance of each feature in predicting price using the MLR.

The importance of each feature in predicting price was also computed using permutation importance, which shuffled the values of features 15 times. The model of a smartphone was found to be most important, which makes sense, since smartphones are often priced according to their model. An iPhone 16 will be the same as another iPhone 16, but differently priced to an iPhone 15. However, because of the sheer number of models included in this dataset, the effect of each is not easily gleaned. Brand is the next most important feature, which can be explained by how companies compete on price, and as such have widely different pricing strategies.

A graph with a number of text

Description automatically generated with medium confidence

Figure 12: Coefficient values of smartphone brands.

Since brand was the most important feature, I further broke its importance up into its individual components. In Figure 12, the coefficient values of each brand were mapped after being divided by their standard deviation to ensure all values were standardized. Cubot was found to have a significant effect on price, meaning that buying a Cubot phone would lower your cost substantially. In contrast, choosing to buy an OPPO phone would increase the price you pay significantly.

## Decision Tree Regressor Model

A decision tree regressor was chosen to model non-linear relationships in the dataset. Because decision tree models split the data using feature combinations, they capture interactions between features well. This is important to this dataset, since it is the combination of features in each smartphone that affects a consumer’s decision to buy it, which ultimately drives the pricing decisions of companies.

A graph with blue dots and a line

Description automatically generated

Figure 13: Decision tree depth plotted against the MSE of test and train data.

To find the optimal maximum tree depth to use for the model, the MSE scores of train and test data was found for each max tree depth and plotted in Figure 13. A depth of 6 was used as this had the lowest test MSE. Using this max depth, a pipeline was created to feed train and test data through.

A screenshot of a computer program

Description automatically generated

Figure 14: Pipeline for the decision tree regressor.

Using the decision tree regressor, the **MSE for training data was 23,584.41**, which was much higher than the MLR’s score of 9,954.23. However, its **MSE score for test data was 23,462.57**, which indicated a much better performance than the MLR’s 46,855.46 score. Since the MSE score for both training and test data was similar, the decision tree model performed consistently across both sets. This indicates it is generalizing well to the unseen test data, which can be attributed to finding the optimal max tree depth, which has prevented overfitting/underfitting.

A screenshot of a phone

Description automatically generated

Figure 15: Importance of each feature in predicting price using the decision tree regressor.

As in Figure 15, RAM was the most important indicator of smartphone prices, according to the decision tree regressor. This was followed by smartphone model, which was most important to the MLR, then brand. These differences can be attributed to how these models determine feature importance differently. Whereas the MLR used the coefficients of each feature to assign importance, the decision tree based this on how homogenously the feature split the data.

## Random Forest Regressor Model

Since my decision tree regressor performed well, I decided to use a random forest regressor to see if this would improve performance through multiple decision trees that combine to make its prediction. Random forests can also handle interactions between features well, making it appropriate for this dataset.

To determine the best parameter for the model, a grid search was performed using 5 cross validation folds and scored according to negative MSE. This scoring method was chosen because minimizing prediction error is the metric by which I am assessing these models. It was found that the optimal max depth was 10, and number of estimators was 100.

A screenshot of a computer

Description automatically generated

*Figure 16: GridSearchCV pipeline for random forest regressor.*

It was found that using the random forest regressor led to an **MSE of 14,578.30 for training data**. **For test data, the MSE score was 31,237.49.** The forest model performed better on training data compared to the decision tree, and the MLR still had the lowest training MSE score by far. On test data, however, the forest model performed worse than the decision tree model. The low train and high test MSE scores also indicate inconsistent performance and overfitting.

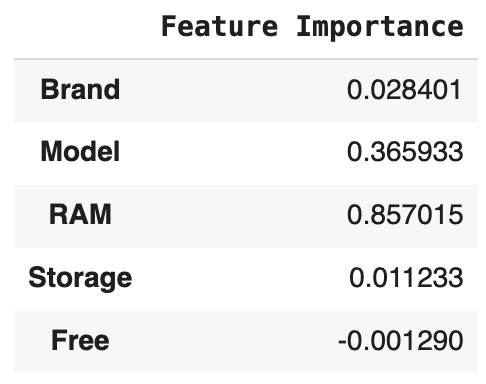


Figure 16: Importance of each feature in predicting price using the forest tree regressor.

Like the decision tree regressor, the forest model ranked RAM as the most important feature, followed by model, brand, then storage. However, it rated whether the smartphone had a free cellphone plan with negative feature importance. This is likely attributed to how only 2.4% of phones in the dataset do not have a free plan, which lowers the relevance of this feature.

# Concluding Discussion

## Summary of Findings

It was found that the decision tree regressor performed best with the lowest testing MSE score of 23,462.57. This was followed by the forest tree regressor, then the multiple linear regression. Nevertheless, all models were significant improvements from the baseline, revealing that utilizing other features in a dataset’s prediction yields more accurate results.

### Key Results

1. **Good Generalization in Decision Tree Model**: The decision tree regressor not only had the lowest test MSE score but was also most consistent among its train/test scores. This shows the importance of finding the optimal max tree depth to fit the model on, which will prevent both overfitting and underfitting – a difficulty that affected the random forest regressor. This additional preparation allowed for the model to generalize well when moving from the train to test data.
2. **The Importance of RAM in Pricing:** According to both the decision tree and random forest model, RAM was the most important feature, with a score of 0.96 and 0.86 respectively. This means that it was the feature most frequently used to split the tree and that it improved the performance of the model most. Model and brand followed in importance for these two models. This was surprising to me, since RAM is not often a feature that consumers take note of, based on personal experience. This demonstrates that consumer preference is not the primary factor driving pricing decisions, and that a smartphone’s physical attributes are far more important in predicting price than non-physical ones.

## Next Steps

By gathering additional data on the smartphones in this dataset, this model could enhance its predictive capabilities.

### Physical Attributes

* **Camera capabilities:** the quality of photos taken on a smartphone is becoming increasingly valuable to consumers, so including this specification would likely prove useful.
* **Battery life:** longer battery life is another key feature consumers value when choosing a smartphone.
* **Display size:** this would show if larger or smaller phones are priced higher.

### Non-physical Attributes

* **Perceived brand value:** consumers could be asked to rate their value of a brand using a survey, which may have interesting interactions with smartphone prices.
* **Model age:** the newness of a phone could have a relationship with how it is priced. It may also be interesting to look at a singular phone model and track its price over time.

# Appendix

## Appendix A: Early Iteration of the MLR

A screenshot of a computer

Description automatically generated

Importance of each feature in predicting price.

Color had a negative feature importance value of -0.000432, so it was removed from the final MLR model. This improved the test data performance from an MSE of 47,055.18 to 46,855.46.