Capstone Project

Anastasiia
Reznichenko Data Science Nanodegree
September 18, 2023

Definition

Project Overview

Today, it’s hard to imagine life without a mobile phone, right? But when you dive into the world of buying a new phone, it can be a real head-spinner. There are so many options out there, each with its own set of features, and the price range is all over the place.

In this project, we aim to investigate how smartphone prices are determined. We want to identify which features contribute to higher prices and which ones are unrelated to the price. Our goal is to construct a model that can predict smartphone prices based on these features. Ultimately, this research could empower consumers to make more informed decisions when purchasing a phone. It highlights the idea that sometimes we pay for a renowned brand name rather than the internal specifications of the phone.

Problem Statement

The primary objective is to comprehend the key features influencing smartphone prices. Initially, we intend to explore the potential of utilizing the inflation rate to forecast future phone prices. Concurrently, we will develop a predictive model employing RainForestRegressor to estimate prices based on chosen features.

To answer the question if we can predict prices considering only inflation rate, we plan to compute the average price for various brands in 2017 and project their estimated prices in 2023 considering the inflation rate.We would compare real and estimated prices.

After that we would build a model employing RainForestRegressor.Wa are using RandomForestRegressor because it is a non-linear model, allowing us to uncover both linear and non-linear correlations between features and prices. It should give us an opportunity to predict prices better.

Metrics

To evaluate results we are using common metrics for RainForestRegressor:

- Mean Absolute Error (MAE): This measures the average absolute difference between the predicted and actual values. It provides a straightforward interpretation of the model's error.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} \left| y_i - \hat{y}_i \right|$$

- Mean Squared Error (MSE): MSE calculates the average of the squared differences between predicted and actual values. It gives more weight to larger errors and can be useful for identifying outliers.

$$\text{MSE} = \overset{\text{Mean}}{\boxed{\frac{1}{n}\sum_{i=1}^{n}}}(\overset{\text{Error}}{\boxed{Y_i - \hat{Y}_i}})^{\overset{\text{Squared}}{\boxed{2}}}$$

These metrics were used to evaluate our model and to improve it.

Datasets and Inputs

We are using open source data cleaned_all_phones.csv that we uploaded from https://www.kaggle.com/datasets. It this database we have information about phone names, brands,
type of operating system, size of the phone, resolution, battery model and type, RAM,year of release,weight,storage capacity,types of video formats that are supported,and price.
Also, we took inflation rate from open official public source https://data.bls.gov/cgi-bin/cpicalc.pl

We will incorporate all these features into our model, with the exception of phone names. Additionally, we will leverage the resolution feature to determine the megapixels, as it is a more familiar and comprehensible metric for people.

Solution Statement

Our solution involves a two-fold approach to understanding and predicting smartphone prices:

Inflation Rate Analysis: We aim to investigate the influence of the inflation rate on future phone prices. To address this, we will calculate the average price for various smartphone brands in 2017 and project their estimated prices for 2023, taking into account the inflation rate. We will then compare these projected prices with actual prices to assess the predictive power of the inflation rate alone.

Predictive Model Development: In addition to analyzing the inflation rate, we will construct a predictive model using RandomForestRegressor. This choice is based on the model's non-linear nature, which allows us to capture both linear and non-linear relationships between various features and smartphone prices. Our dataset, sourced from the cleaned_all_phones.csv file, contains a range of features such as brand, operating system, phone size, resolution, battery specifications, RAM, release year, weight, storage capacity, supported video formats, and price. We will use all of these features, except phone names, in our model. Furthermore, we will utilize the resolution feature to infer megapixels, as it is a more familiar metric for most people.

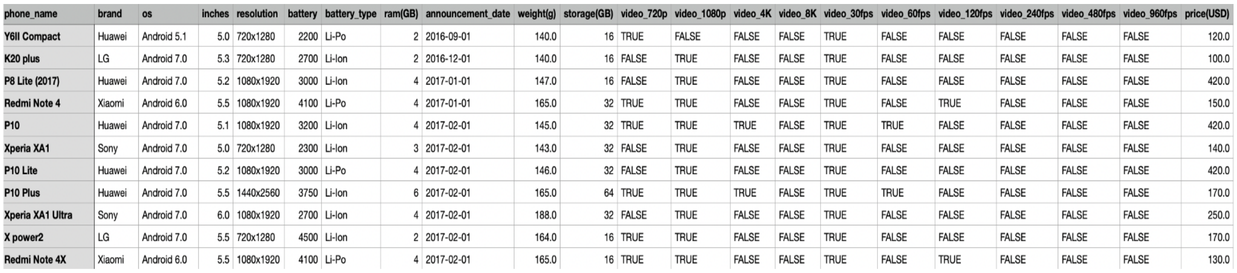
By combining these two approaches, we aim to gain a comprehensive understanding of the factors influencing smartphone prices and develop an accurate predictive model that can assist consumers and stakeholders in making informed decisions about smartphone purchases.

Exploratory Data Analysis

Data exploration and Visualization

We collected open-source data on various phone brands and models released between 2017 and 2023.

We have gathered data on model names, brands, and various features such as battery type, operating system (iOS), phone size, resolution, battery specifications, RAM, release year, weight, storage capacity, supported video formats, and price.

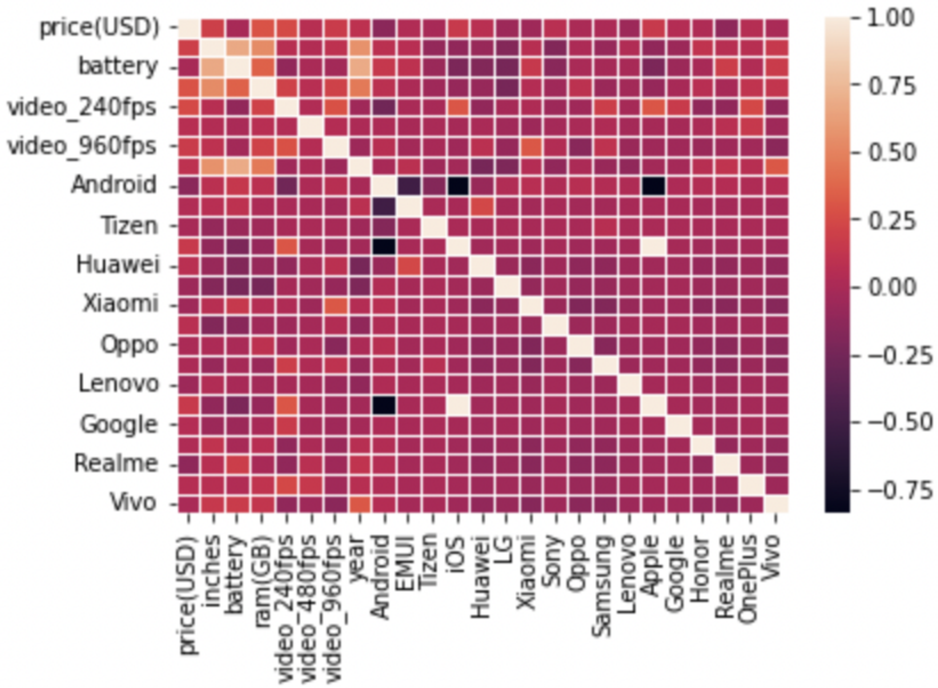


phone_name	brand	os	inches	resolution	battery	battery_type	ram(GB)	announcement_date	weight(g)	storage(GB)	video_720p	video_1080p	video_4K	video_8K	video_30fps	video_60fps	video_120fps	video_240fps	video_480fps	video_960fps	price(USD)
Y6II Compact	Huawei	Android 5.1	5.0	720x1280	2200	Li-Po	2	2016-09-01	140.0	16	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	120.0
K20 plus	LG	Android 7.0	5.3	720x1280	2700	Li-Ion	2	2016-12-01	140.0	16	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	100.0
P8 Lite (2017)	Huawei	Android 7.0	5.2	1080x1920	3000	Li-Ion	4	2017-01-01	147.0	16	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	420.0
Redmi Note 4	Xiaomi	Android 6.0	5.5	1080x1920	4100	Li-Po	4	2017-01-01	165.0	32	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	150.0
P10	Huawei	Android 7.0	5.1	1080x1920	3200	Li-Ion	4	2017-02-01	145.0	32	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	420.0
Xperia XA1	Sony	Android 7.0	5.0	720x1280	2300	Li-Ion	3	2017-02-01	143.0	32	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	140.0
P10 Lite	Huawei	Android 7.0	5.2	1080x1920	3000	Li-Po	4	2017-02-01	146.0	32	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	420.0
P10 Plus	Huawei	Android 7.0	5.5	1440x2560	3750	Li-Ion	6	2017-02-01	165.0	64	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	170.0
Xperia XA1 Ultra	Sony	Android 7.0	6.0	1080x1920	2700	Li-Ion	4	2017-02-01	188.0	32	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	250.0
X power2	LG	Android 7.0	5.5	720x1280	4500	Li-Ion	2	2017-02-01	164.0	16	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	170.0
Redmi Note 4X	Xiaomi	Android 6.0	5.5	1080x1920	4100	Li-Po	4	2017-02-01	165.0	16	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	130.0

1.**First case**.To begin, our initial step is to estimate prices for 2023 based on the data from 2017, taking inflation into account. If we observe that the estimated prices closely align with the actual prices, this suggests that a single inflation factor can effectively predict future phone prices. This analysis will help us gauge the predictive power of inflation in the smartphone market.
It’s clear from the analysis that prices have increased significantly more than what can be attributed to inflation alone. This suggests that predicting phone prices solely based on inflation is not sufficient.

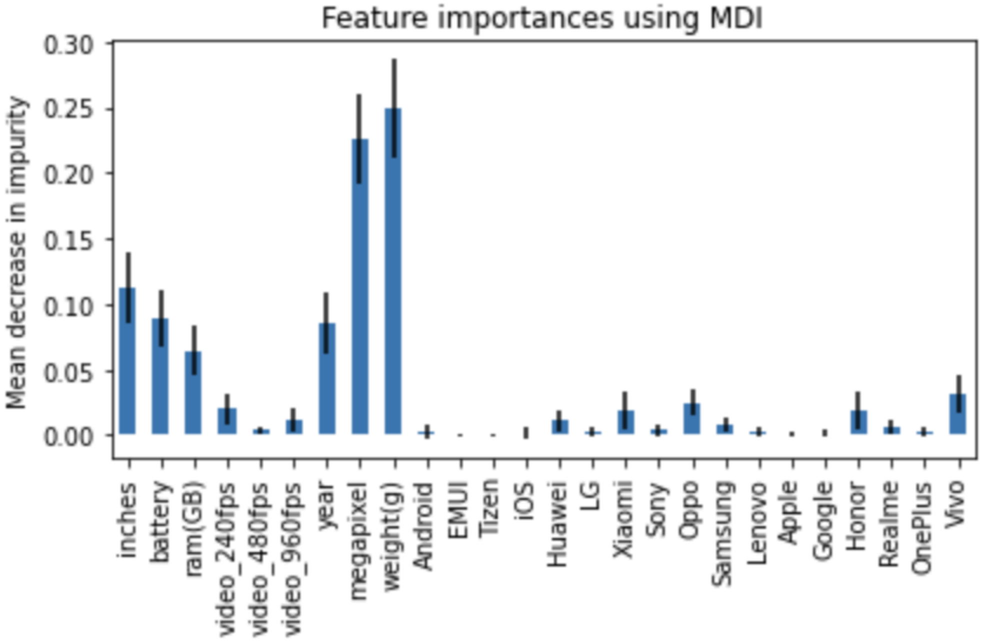
2.Second case.Building a model.If we are not discussing iPhones, where the next model name typically follows a numerical sequence by adding 1 to the previous model, the naming convention for phones can be quite unpredictable. Therefore, we won't be able to discern any discernible pattern, and as a result, we won't include it in our analysis.

To understand if phone features like size, model,brand and etc. have any correlation with the price, we can calculate the correlation matrix and visualize it using a heatmap. This will help us identify which features are most strongly correlated with the price of the phone.



We cannot definitively conclude that a single feature solely determines the pricing of a phone. However, we can observe that certain features have a stronger correlation with the price compared to others.

Certainly, using RandomForestRegressor as a non-linear model allows us to calculate feature importances. These important values are instrumental in gaining insights into which features exert the strongest influence on smartphone prices. This analysis aids in identifying the key factors that drive price variations in the smartphone market.



We can observe that factors like the size of the phone, battery capacity, RAM,,the year of release,megapixels and weight exhibit the strongest correlations with phone prices. On the other hand, the type of operating system (OS) doesn’t seem to contribute significantly and can be removed to enhance our model.

Algorytm and Technics

For creating the inflation table, we employed a straightforward algebraic approach. We initially grouped all the prices by brand and name, computing the average price for each. Subsequently, we took the price in 2017 and multiplied it by the inflation rate from 2017 to 2023. This inflation rate calculation was based on the formula: 1 +percent inflation/100.

This method allowed us to estimate prices for 2023, providing a practical and comprehensible way to incorporate inflation dynamics into our analysis.

In the second step of our project, our objective was to construct a predictive model capable of estimating smartphone prices based on various features. When we examined the correlation matrix, we observed that there wasn't a strong linear correlation between the target variable (price) and the other features. Due to this lack of strong linear relationships, we made the informed choice to utilize the RandomForestRegressor as our model.

For this model, we designated "price" as our target column and included all other columns (except for "phone name") as our feature columns. RandomForestRegressor is a powerful choice because it is a non-linear model capable of capturing both linear and non-linear relationships between features and the target variable. This flexibility makes it well-suited for our prediction task, where linear correlations may not fully represent the complexity of the data.

Hyperparameter tuning:

To improve our model we can tune such parameters:

- Number of Trees.We can change the number of decision trees in the ensemble
- Maximum Depth of Trees
- Minimum samples required to split a node
- Minimum samples required at each leaf node

We utilized the `GridSearchCV` function to systematically search for the optimal hyperparameters for our model.

BenchMark

We created a simple baseline model that predicts smartphone prices based solely on the average price of all smartphones in the training dataset.For a common evaluation metric we took MAE (Mean Absolute Value Error). The goal is to see if our model outperforms the simple baseline.

While a simple baseline model has MAE: 156.78 our model that was built using RainForestRegressor after all improvements has MAE: 124.03 which is significantly smaller.

That means that our model provides better results.

Methodology

Data Preprocessing

All process to prepare data for the model was done in notebook “Phone.ipynb”.Next steps were performed:

1.Converting Boolean Variables: All boolean variables were transformed into binary values (0 or 1).

2.Creating 'Width' and 'Length' Columns: A 'Width' and 'Length' column was created by splitting the 'Resolution' column.

3.Calculating 'Megapixels': A 'Megapixels' column was generated using the 'Width' and 'Length' columns.

4.Extracting 'Year': A 'Year' column was derived by converting the 'Announcement Date' column to a date type and extracting the year from it.

5.Fixing Misspellings: Corrections were made for any misspelled values or inconsistencies in the dataset.

6.Creating 'os_type' and 'os_model' Columns: The 'os'' column was split to create separate 'os_type' and 'os_model' columns.

7.Dummy Columns: Dummy columns were created for the 'brand' and 'os_type' variables.

8.Aggregating Data: A new table was generated by grouping phone brands and years and calculating the average price for each group.

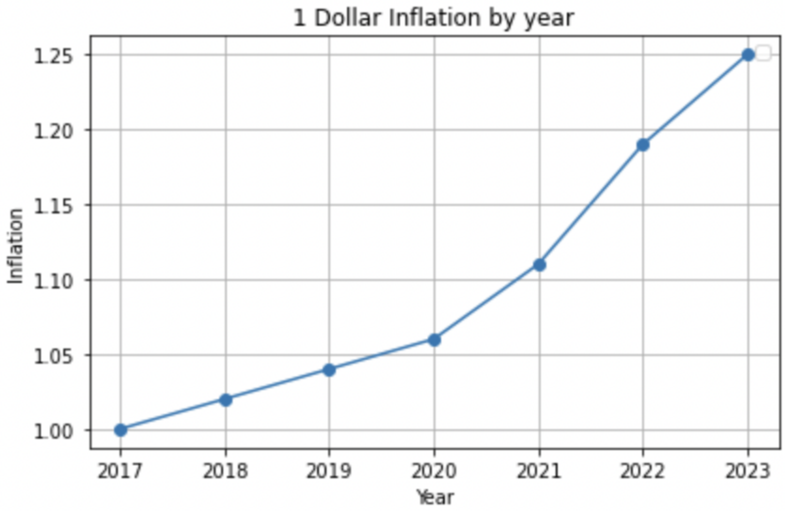
9. Handling Missing Data: Entire rows for the year 2016, where more than 80% of the information was missing, were dropped. Missing values were filled using forward and backward filling methods

Implementation

The implementation phase can be divided into the following stages:

- Prediction Based on Inflation Rate: Initially, predictions were made based on the inflation rate. The average price for various smartphone brands in 2017 was calculated, and estimated prices for 2023 were projected, considering the inflation rate. Real and estimated prices were compared.
- Building a Model: Subsequently, a predictive model was constructed. The RandomForestRegressor model was chosen due to its ability to capture both linear and non-linear correlations between features and prices. All relevant features, except 'Phone Name,' were used.
- Model Improvement with Grid Search: To enhance the model's performance, hyperparameter tuning was performed using GridSearchCV. This optimization technique systematically searched for the best combination of hyperparameters to improve predictions.

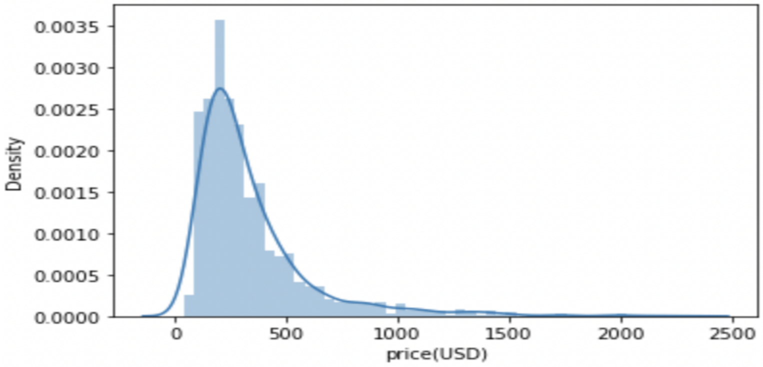
During the first stage we did Prediction Based on Inflation Rate.First we investigate the inflation since 2017.



Using the data for 2017, we calculated the inflation rate specific to our dataset. Subsequently, we proceeded to compare the predicted prices with the actual prices, revealing significant disparities between the two.

This revised statement provides a more concise and clear description of the steps taken to calculate inflation and the observed differences between predicted and actual prices.

In the second stage, we constructed a predictive model and trained it using the RandomForestRegressor algorithm. To enhance the model's performance, an initial step involved identifying and addressing potential outliers. To detect outliers, we employed sns.distplot to visualize the 'Price' column, with the aim of identifying extremely high or low-priced phones that might influence our model's accuracy.



This revised description provides a clearer account of how outlier detection was carried out to improve the model's quality.

And the last step was to improve the model to enhance the model's performance.For that purpose we used GridSearchCV.

The combination of these stages helped in preparing the data, building an initial model, and refining it to achieve better predictive capabilities.

These improvements provide a clearer and more organized description of your data preprocessing and implementation steps.

Results

Model Evaluation and Validation

In our model evaluation process, we opted for two widely recognized evaluation metrics in the field of price prediction: Mean Absolute Error (MAE) and Mean Squared Error (MSE). These metrics provide valuable insights into the accuracy and performance of our model, making them suitable choices for assessing its predictive capabilities.

To fine-tune our model and ensure its optimal performance, we harnessed the power of GridSearchCV. This technique systematically explored a range of hyperparameters, allowing us to pinpoint the best combination for our RandomForestRegressor model. This meticulous hyperparameter tuning contributed to the model's ability to make more accurate price predictions.

To validate our model's effectiveness and assess its generalization to unseen data, we meticulously split our dataset into two distinct subsets: a training set comprising 80% of the data and a test set encompassing the remaining 20%. By training the model on the training data, we enabled it to learn from the patterns and relationships within that portion of the dataset. Subsequently, we evaluated the model's performance by applying it to the test set, which simulates real-world predictions.

The application of GridSearchCV for hyperparameter tuning was an integral part of our model improvement strategy, as it enabled us to identify the most suitable hyperparameters for our RandomForestRegressor model. This optimization process enhanced the model's ability to make precise price predictions.

In addition to these rigorous evaluation and tuning procedures, we visualized the predicted prices and compared them to the actual test prices. This visualization step provided us with a tangible understanding of our model's predictive capabilities, allowing us to assess how effectively it could forecast smartphone prices.

This refined description highlights the critical role of evaluation metrics, hyperparameter tuning, dataset splitting, and visualization in our model evaluation and improvement process.

Conclusion

In conclusion, our predictive model for smartphone prices demonstrated promising performance based on the evaluation metrics:

- Mean Squared Error (MSE): 35,159.08
- Model Mean Absolute Error (MAE): 124.75

These metrics provide valuable insights into the model's predictive accuracy and its ability to estimate smartphone prices. While the MSE quantifies the average squared difference between predicted and actual prices, the MAE measures the average absolute difference. In both cases, lower values indicate better predictive performance.

Our model's MAE of 124.75 suggests that, on average, it predicts smartphone prices with an error of approximately $124.75. While this error is relatively small considering the price range of smartphones, further refinements and feature engineering may help improve prediction accuracy.

In summary, our model lays a solid foundation for predicting smartphone prices based on various features. It has the potential to assist consumers in making informed decisions when purchasing smartphones, taking into account factors that influence pricing. Future work could involve fine-tuning the model further, exploring additional features, and expanding the dataset to enhance prediction accuracy.

With a Mean Squared Error (MSE) of 35,159.08 and a Model Mean Absolute Error (MAE) of 124.75 for your RandomForestRegressor model, there is certainly room for improvement. Exploring alternative models and fine-tuning the existing one can potentially lead to enhanced predictive performance.

Here are some strategic steps to consider for refining themodel and optimizing its predictive accuracy:

Feature Set Enhancement: we nned to conduct a comprehensive review of the current feature set. Explore the possibility of incorporating additional pertinent features that might contribute to a more accurate prediction.

- Hyperparameter Refinement: Continue the process of fine-tuning the hyperparameters of the RandomForestRegressor model. Employ techniques like Grid Search or Random Search to systematically experiment with various hyperparameter combinations and identify the configuration that minimizes MSE and MAE.

- Ensemble Strategies: Embrace the power of ensemble techniques by combining multiple models. We can consider integrating the RandomForestRegressor model with other regression models, such as Gradient Boosting (e.g., XGBoost, LightGBM), AdaBoost, or even another RandomForest variant with distinct hyperparameters. Ensembling these models can often yield substantial improvements in prediction accuracy.

We nned to be sure to maintain a meticulous record of experiments, capturing the outcomes and insights derived from each iteration. This documentation will serve as a valuable resource for making informed decisions regarding the most effective enhancements to themodel.

Furthermore, consider conducting a thorough error analysis to gain insights into your model's strengths and weaknesses. Scrutinize the nature of predictions, particularly instances where the model exhibits deviations from actual values. This analysis can guide targeted refinements for even better performance.