CBD3334
Data Mining and Analysis

Mobile Price prediction using phone Specifications - Regression

Group Name: Team Danilo

Group Members:

First name	Last Name	Student number
Aanal	Patel	C0910376
Bimal Kumar	Shrestha	C0919385
Danilo	Diaz	C0889539
Ernie	Sumoso	C0881591
Jayachandhran	Saravanan	C010392

Submission date: 16-Apr-2024

Template Prepared by: William Pourmajidi

Updated by Vahid Hadavi Nov 2023

Contents

Abstract 3
Introduction 4
Dataset 5
Methods Exploratory Data Analysis: 6
Visualization 8
Preprocessing 8
Training and model testing 10
Model table (baseline and hyper-parameter tuning) 10
Conclusions 12
Insights: 12
Annexures 13
References 15

Abstract

This project will discuss the various steps involved, from fetching the raw data to comparing the performance of the models. In this activity, the data speaks about mobile phone prices and the respective important specifications (features) of that particular mobile phone. Standard data operations are applied throughout the assignment to address the distribution of the numeric data and handle the outliers. Using feature selection, the top five features are listed, and their importance is compared with each trained model to verify the impact of the selected feature on the predictions. The linear regression model shows decent performance as it is considered the baseline model in this activity; further, the residuals are moderate compared to other models. The regularization and hyperparameter tunings are applied, and the results are discussed and elaborated. Lasso outperformed the used models, with good performance on both the testing and training datasets and a mean absolute error of 4200(Rs) . Random forest and gradient-based boosting models show slight overfitting characteristics. The following report shows a complete overview of the regression-based approach to predicting the mobile phone(smartphone) price.

Key Words: regression, mobile phone, feature selection, linear regression, Lasso regularization (L1 Norm), Mean Square Error, random forest, gradient-based boosting, baseline model and smartphone

Introduction

The usage of mobile phones in this modern era is considered to be one of the basic necessities along with other basic needs. The smartphone brings the whole world into the user's palm, and choosing it requires more detailed research comparisons. Each year, technology grows rapidly and also determines the price of the smartphone with the presence of specifications which include but are not limited to touchscreen, dual SIM, global positioning sensor (GPS), Bluetooth capabilities, processing chip, internal storage and processing memory. In this report, a detailed analysis of smartphone prices is carried out, and the regression model to predict the upcoming mobile phone price is attained. There is still room for improvements in tuning the model's performance and considering more important features which add more weight to the model during the training phone. The following is the task distribution of this project.

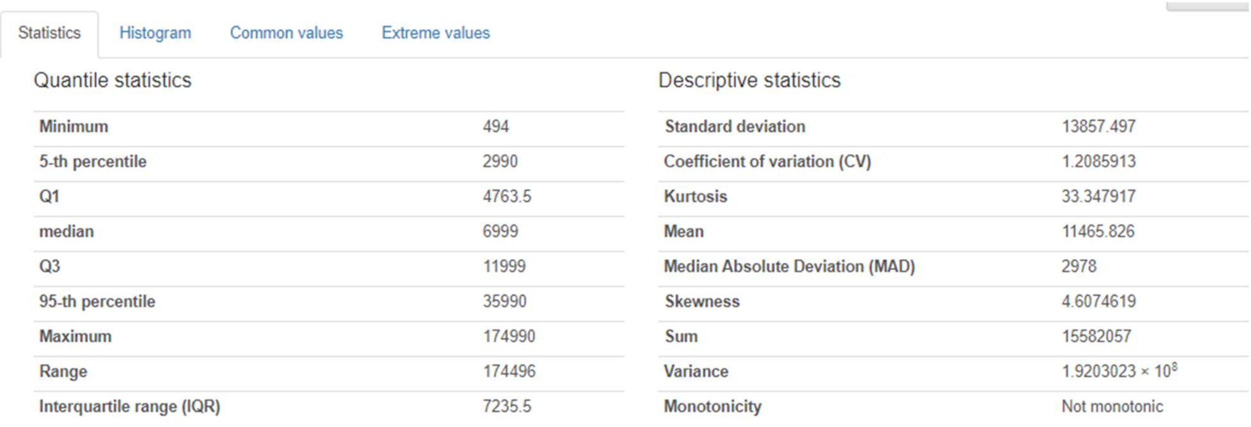
TASK	PURPOSE/DESCRIPTION
Data source	- Finalizing an informative dataset to carry out the mentioned Machine Learning Steps to gain more hands-on experience
Feature set understanding	- 360-degree examination of the dataset to articulate the further analysis
Exploratory Data Analysis	- Handling the data instabilities and uncovering the characteristics of the dataset
Categorical and Numerical Feature processing	- Special attention to the feature types as they impact the performance of the model
Pre-Processing	- Standardization and normalization of the data points to create some meaningful pattern for the model to learn
Model Pipeline	- Significant steps to experiment with various models and their predictions on different types of data combination
Tuning Model parameters	- tweaking and checking the performance of the models based on their potential parameters (e.g. learning rate, iterations)
Results Comparison	- Observation and inference of the performance with the known and unknown dataset

Dataset

The data is fetched from the Kaggle website, where the ultimate data describes the characteristics of most phones in the market. The total size of the data is around 1,400 rows, and 22 features as columns. Further, the dataset terminology is explained below,

Feature Name	Representation
Name	Mobile (smart) phone name
Brand	The brand which entitles the phone
Model	Model name specifies the version of the particular mobile phone (e.g., iPhone13 has mini, Pro, and Promax as model names)
Battery	Power Capacity of the phone mentioned in milli Ampere (hour)
Screen size	diagonal measurement of the screen in inches
Resolution x and y	Represents the screen pixel rate with respect to the X and Y axis
RAM	Processing Power (memory) in Megabytes
Processor	Specification of the processor chip
Internal Storage	The storage capability of the phone is measured in Gigabytes
OS	Operating System details
Rear and Front Camera	Camera resolution capturing specifications
Number of sims	Sim slots available on the phone
Feature set columns	• Bluetooth • Wi-Fi • 3G and 4G • GPS • Touch screen
Price	Selling price (in Indian Rupees)

Pandas Profiling Report:



Quantile statistics

Minimum	494
5-th percentile	2990
Q1	4763.5
median	6999
Q3	11999
95-th percentile	35990
Maximum	174990
Range	174496
Interquartile range (IQR)	7235.5

Descriptive statistics

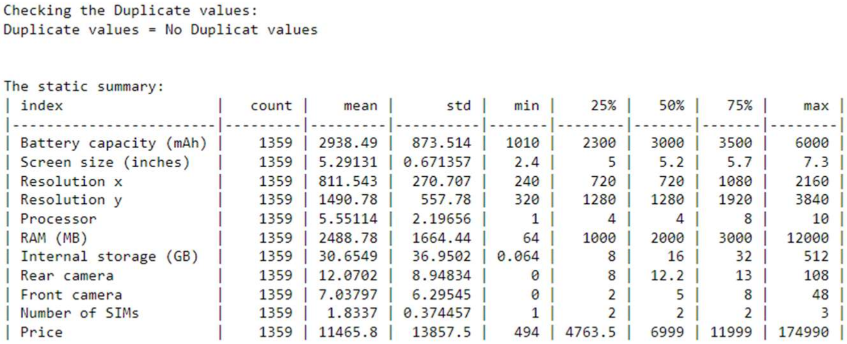
Standard deviation	13857.497
Coefficient of variation (CV)	1.2085913
Kurtosis	33.347917
Mean	11465.826
Median Absolute Deviation (MAD)	2978
Skewness	4.6074619
Sum	15582057
Variance	1.9203023×10^8
Monotonicity	Not monotonic

Using Pandas profiling reports, more statistical-based inferences were attained, and it also helped get more details on the data with clear objectives.

Methods

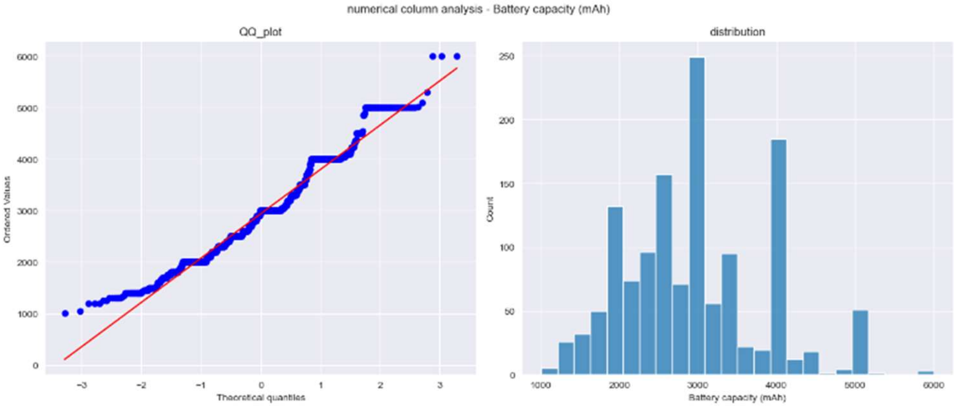
Exploratory Data Analysis:

This section encloses the required data analysis carried out. The raw data is identified with the shape of (1359, 22). The primary data validation of checking the missing and duplicate values is checked, and the data is more likely processed before fetching, so a very negligible amount of values is seen. The extensive analysis of the data, like statistical analysis of the numeric columns using describe method provides more meaningful insights into the data distribution and the important findings present in the data. Further, the data is checked for outlier values. In this activity, the numeric columns are



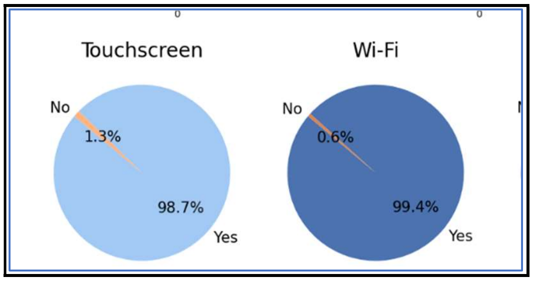
```
Checking the Duplicate values:
Duplicate values = No Duplicat values


The static summary:
| index                  |   count |    mean |      std |    min |    25% |   50% |   75% |    max |
|------------------------|---------|---------|----------|--------|--------|-------|-------|--------|
| Battery capacity (mAh) |    1359 | 2938.49 |  873.514 |   1010 |   2300 |  3000 |  3500 |   6000 |
| Screen size (inches)   |    1359 | 5.29131 | 0.671357 |    2.4 |      5 |   5.2 |   5.7 |    7.3 |
| Resolution x           |    1359 | 811.543 |  270.707 |    240 |    720 |   720 |  1080 |   2160 |
| Resolution y           |    1359 | 1490.78 |   557.78 |    320 |   1280 |  1280 |  1920 |   3840 |
| Processor              |    1359 | 5.55114 |  2.19656 |      1 |      4 |     4 |     8 |     10 |
| RAM (MB)               |    1359 | 2488.78 |  1664.44 |     64 |   1000 |  2000 |  3000 |  12000 |
| Internal storage (GB)  |    1359 | 30.6549 |  36.9502 |  0.064 |      8 |    16 |    32 |    512 |
| Rear camera            |    1359 | 12.0702 |  8.94834 |      0 |      8 |  12.2 |    13 |    108 |
| Front camera           |    1359 | 7.03797 |  6.29545 |      0 |      2 |     5 |     8 |     48 |
| Number of SIMs         |    1359 |  1.8337 | 0.374457 |      1 |      2 |     2 |     2 |      3 |
| Price                  |    1359 | 11465.8 |  13857.5 |    494 | 4763.5 |  6999 | 11999 | 174990 |
```

validated using more than one method to check the presence of the outliers. The reason is to use more than one technique to understand the outliers and develop a proper prevention or handling pipeline. After looking at the descriptive table, the numeric values are plotted using a line plot along with the target variable to check the spikes and longevity of the tail or extreme values.



The five-number summary statistics are used to get the interquartile range (IQR), which is then visualized using boxplots to capture the potential data points. The QQ (prob_plot) method is used to substantiate the theoretical representation of the data points against the actual values.

It clearly explains the distribution and provides a comparative analysis of the projected values. Similarly, the binary data types are analyzed using the pie chart and the Boolean representation ratios are clearly noted. The Frequency chart for the categorical values is displayed for the potential features to interpret its values. The frequency count also provided the variance of the column values, which are highly required for feature selections. Finally, the informative analytics of the features are printed, and the data types are verified.



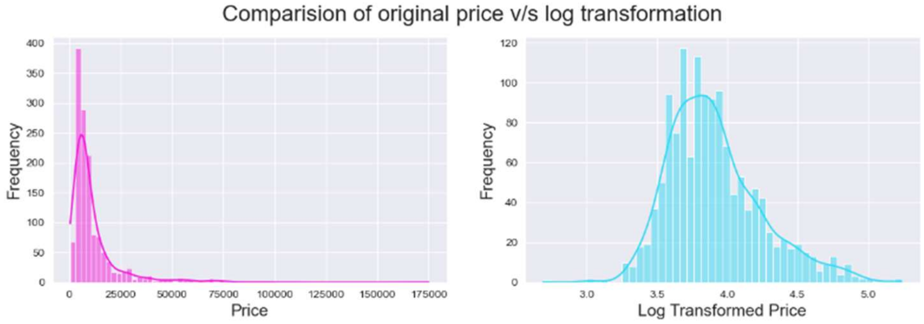
Visualization

Numerous types of visual representation tools are utilized in this project for various analytical reasons. The following table gives an oversight view of the visual graphs used and their purpose.

Boxplots	Capturing potential outliers
Distribution and QQ plot	Detailed analysis of the outliers w.r.t columns & the target variables
Frequency chart	Understand the unique values present in the categorical columns
Pie Chart	Ratio presentation of the binary values
Heatmap	Correlation of features

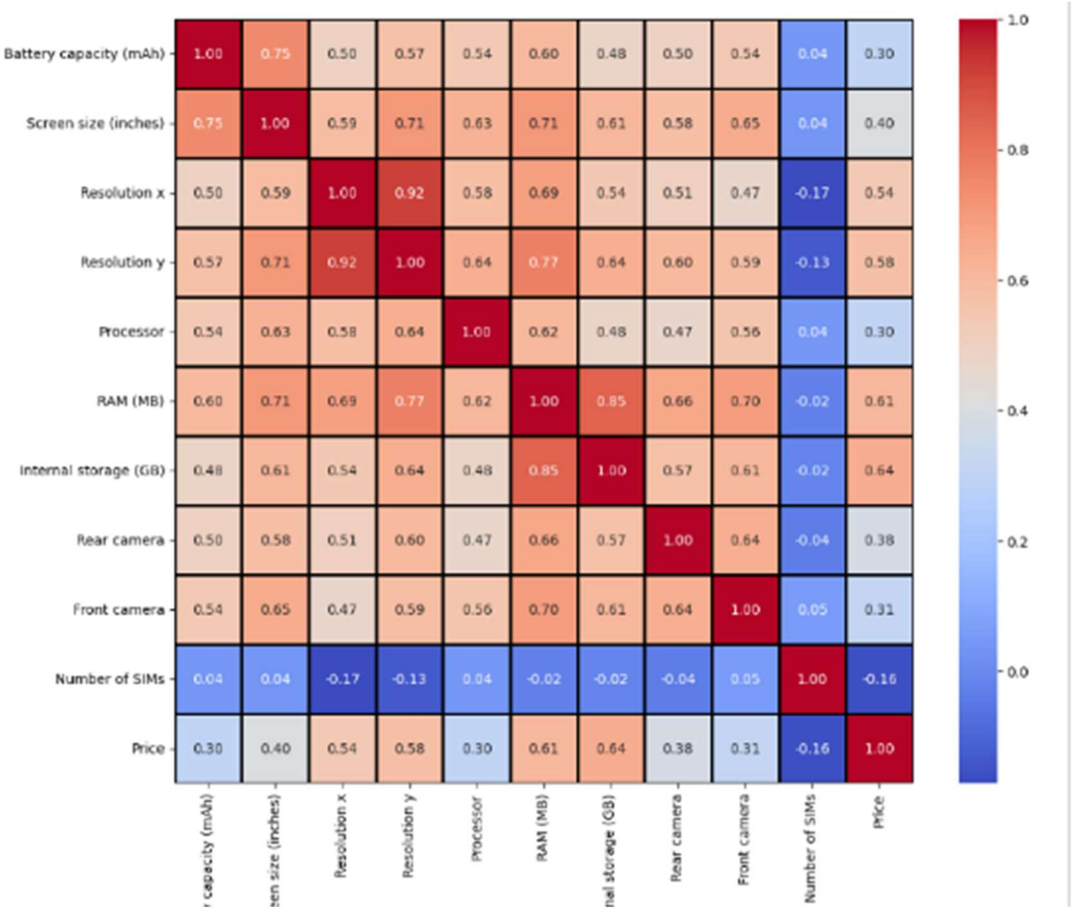
Preprocessing

The pre-processing steps include handling outliers present in the numeric columns, validation of the categorical and numeric values, target variable correlations and potential multi-collinearity handlings. The categorical features are processed and converted to numeric representation using the label encoding technique. With experimentation of all the available approaches to handle outliers, this dataset is more suitable for applying logarithmic transformation to preserve the originality of the data and avoid unwanted bias in the cleaned data. After the numeric values are processed, the outlier is quite less significant in impacting the model’s performance. The visual representation of the numeric values and


Comparision of original price v/s log transformation

their distribution is plotted using a histogram method with density estimation. It emphasizes the distribution is normalized with a standard deviation of 1 and mean value of 0.

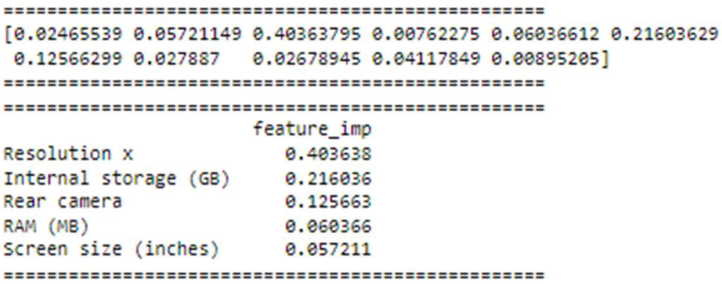
In addition to handling the outliers, feature removal is applied to ignore the less significant columns like Names, models and brands from training. The multilinear data features like the resolution of the Y axis, screen size and the front camera are avoided. Finally, feature engineering is applied to the five categorical columns, creating the latest tech flag column. It is defined as a Boolean data type with "1" representing the presence of all the features and "0" resembling the absence of even one feature. Initial log transformation was applied to the target variable, and the results were not good, so the transformation was applied to all the numeric values.



Results

Training and model testing

The processing pipeline is constructed with one transformer step and the regression model step. Then, the experiment showed better results on the combination of pre-processed categorical value is allowed through the "passthrough" flag, thereby reducing the steps involved in the transformation. The numeric values are scaled using the standard scaler method with a mean value of 0 and a standard deviation of unit 1. Prior to this, the features are mapped to the SelectKBest method to get the best 5 features of the given dataset. The linear regression model is considered as the baseline model, and the train_test set values are configured in a 70:30 ratio with selected features. The training pipeline is optimized in a way that fits the data to the model and gets the prediction values for both the training and testing datasets.

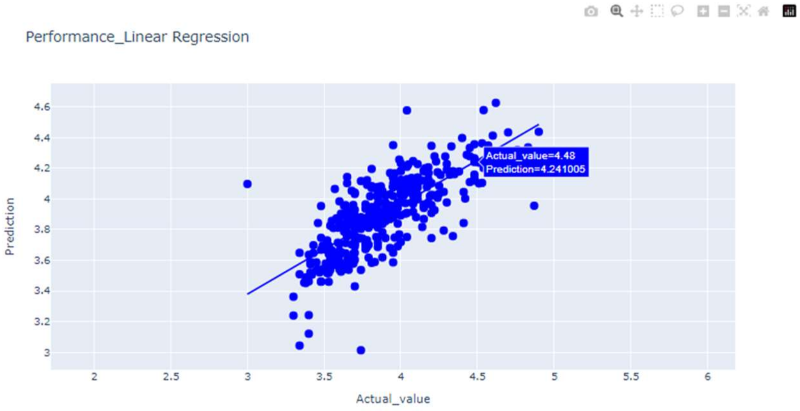


```
==================================================
[0.02465539 0.05721149 0.40363795 0.00762275 0.06036612 0.21603629
 0.12566299 0.027887   0.02678945 0.04117849 0.00895205]
==================================================
==================================================
                       feature_imp
Resolution x              0.403638
Internal storage (GB)     0.216036
Rear camera               0.125663
RAM (MB)                  0.060366
Screen size (inches)      0.057211
==================================================
```

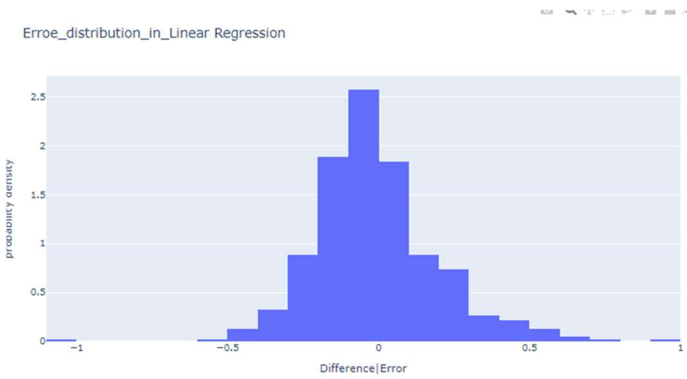
Model table (baseline and hyper-parameter tuning)

Model	Remarks
Linear Regression	Decent Performance on both testing and training data
Random forest regressor	High overfitting
Gradient Boosting regressor	Slight overfitting
Random forest (hyperparameters)	Slight overfitting
Gradient Boosting (hyperparameters)	Overfitting is seen
Ridge Regularization	Near moderate performance
Lasso Regularization	Best performance on both testing and training

The model is evaluated using standard metric systems like R2, Mean Square Error and Mean Absolute Error scores. The values are saved and formulated as a comparison table to investigate the results of various combinations of datasets. Post-training, the importance of the features is carried out to get the most important features for each model. The top 5 weighed features are compared with the Kbest 5 features selected during the pre-training module. Surprisingly, the feature sets are closely associated, but the weighting is varied, with collinearity detected. The scatter plot is implied here to show the actual and predicted data points visually, and the regression lines are plotted along to give a better understanding.



The residuals of each model are plotted with a density curve, and the histogram representation is displayed for various references. Using Plotly, an interactive graphical method, the evaluation metrics are shown, and the feature of closely examining the values is made possible. Thus, the project comprises a detailed approach toward a regression-based prediction problem.

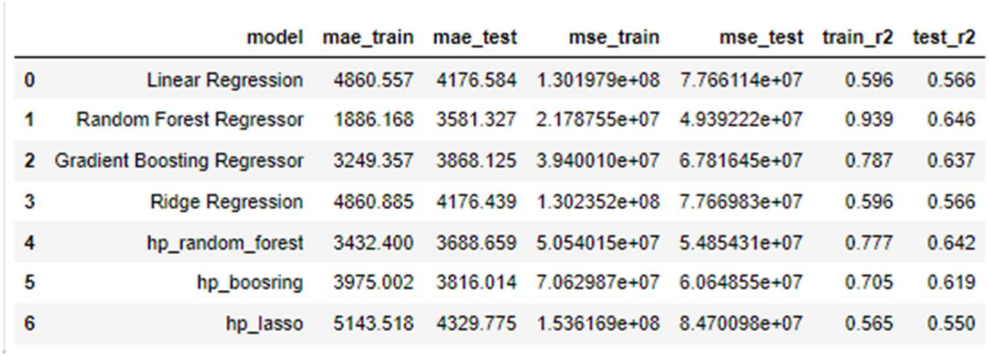


Business Case and Future Scope:

In the future, this project will support numerous price predictions, and further improvements will be helpful in creating a trend analysis of various smart gadgets. It also brings more value to the distributors and the consumers by allowing them to compare the price authentically and effectively instead of paying wholesome, unverified system amounts in the local region. In addition, it will be useful for all kinds of people abiding by the DEI principles

Conclusions

The project analysis and the detailed notebook can be viewed in the hosted remote repository, which contains the dataset and the analysis that was carried out. It also comprises of the final version of the notebook, which discusses the performance of the models along with the pandas profiling. It also provided a clear picture to the team on how to proceed with real-world data and further apply this project in data mining aspects. A neural network-based model and incorporation of live data are planned for the next phase.



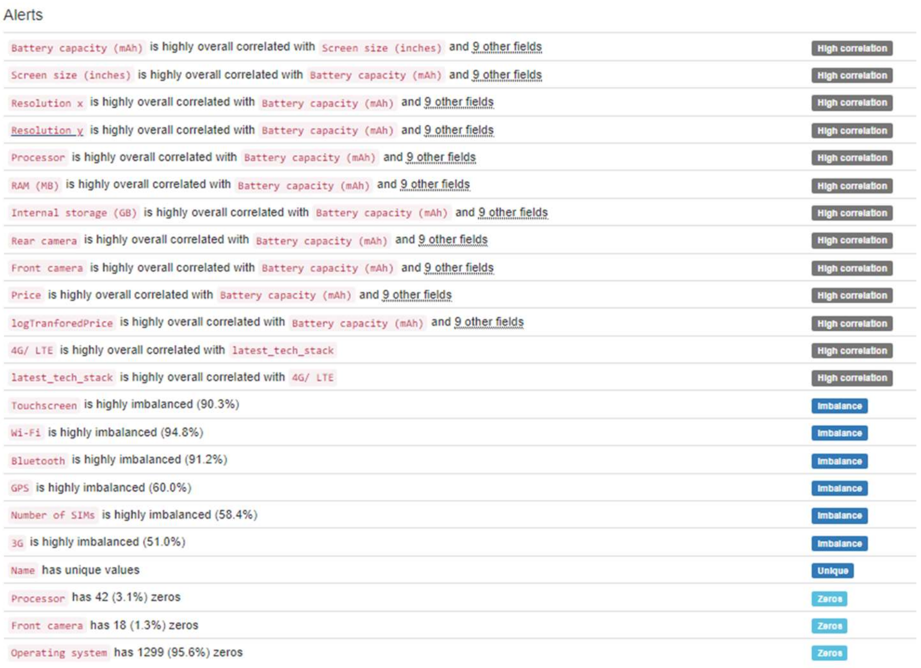
	model	mae_train	mae_test	mse_train	mse_test	train_r2	test_r2
0	Linear Regression	4860.557	4176.584	1.301979e+08	7.766114e+07	0.596	0.566
1	Random Forest Regressor	1886.168	3581.327	2.178755e+07	4.939222e+07	0.939	0.646
2	Gradient Boosting Regressor	3249.357	3868.125	3.940010e+07	6.781645e+07	0.787	0.637
3	Ridge Regression	4860.885	4176.439	1.302352e+08	7.766983e+07	0.596	0.566
4	hp_random_forest	3432.400	3688.659	5.054015e+07	5.485431e+07	0.777	0.642
5	hp_boosring	3975.002	3816.014	7.062987e+07	6.064855e+07	0.705	0.619
6	hp_lasso	5143.518	4329.775	1.536169e+08	8.470098e+07	0.565	0.550

Insights:

- Linear regression shows decent performance in both the training and testing (unknown) dataset; Random Forest and Gradient boosting are overfitting with the training data
- The most important features mostly repeat themselves among our models
- The linear regression model utilizes the lasso regularisation technique, along with various alpha parameters. The best params give the best model for this project activity

Annexures

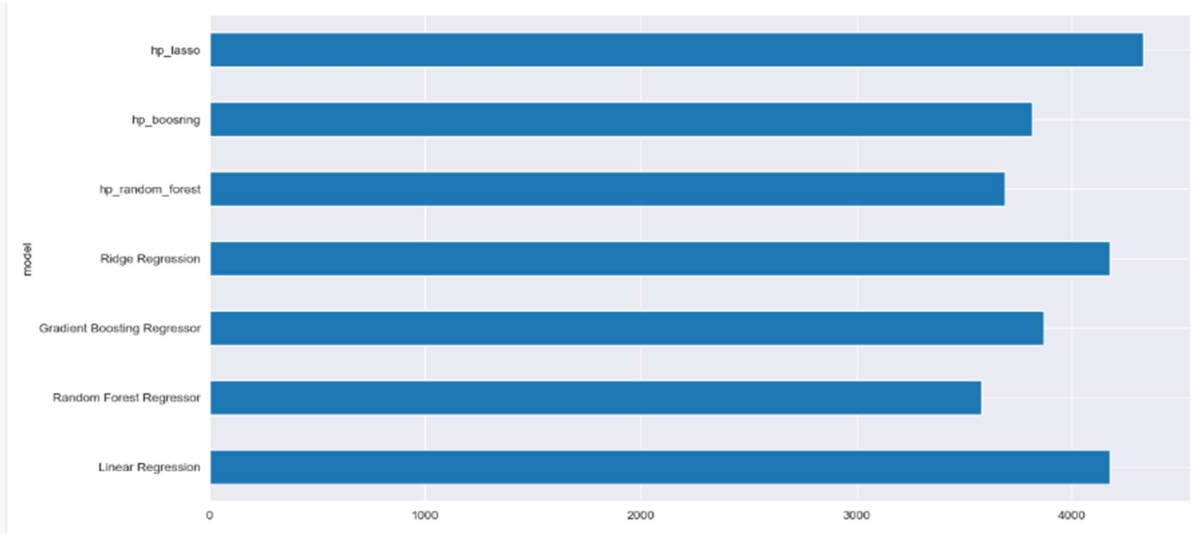
Annexure I: Pandas Profiling

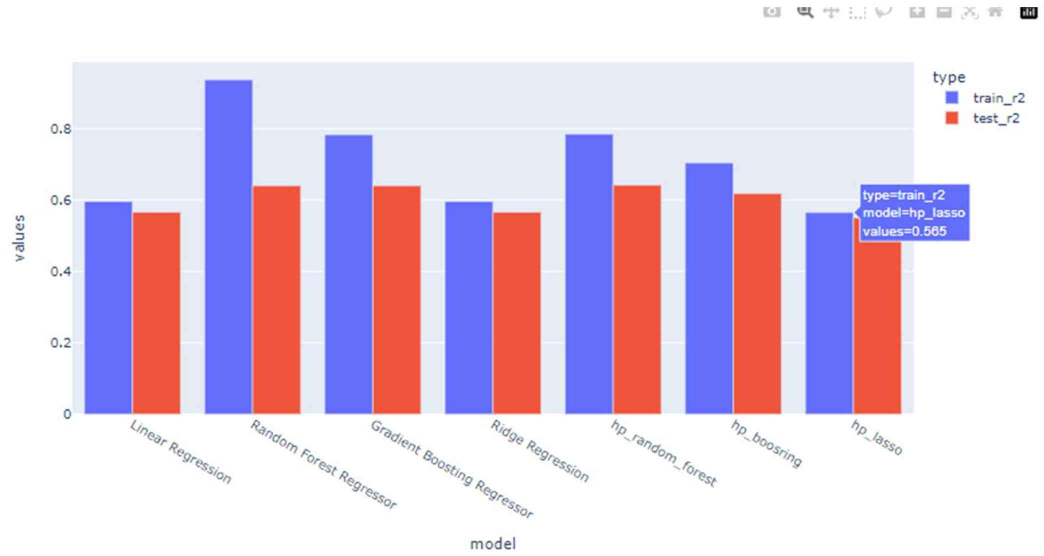


Alerts

`Battery capacity (mAh)` is highly overall correlated with `Screen size (inches)` and 9 other fields	High correlation
`Screen size (inches)` is highly overall correlated with `Battery capacity (mAh)` and 9 other fields	High correlation
`Resolution x` is highly overall correlated with `Battery capacity (mAh)` and 9 other fields	High correlation
`Resolution y` is highly overall correlated with `Battery capacity (mAh)` and 9 other fields	High correlation
`Processor` is highly overall correlated with `Battery capacity (mAh)` and 9 other fields	High correlation
`RAM (MB)` is highly overall correlated with `Battery capacity (mAh)` and 9 other fields	High correlation
`Internal storage (GB)` is highly overall correlated with `Battery capacity (mAh)` and 9 other fields	High correlation
`Rear camera` is highly overall correlated with `Battery capacity (mAh)` and 9 other fields	High correlation
`Front camera` is highly overall correlated with `Battery capacity (mAh)` and 9 other fields	High correlation
`Price` is highly overall correlated with `Battery capacity (mAh)` and 9 other fields	High correlation
`logTranforedPrice` is highly overall correlated with `Battery capacity (mAh)` and 9 other fields	High correlation
`4G/ LTE` is highly overall correlated with `latest_tech_stack`	High correlation
`latest_tech_stack` is highly overall correlated with `4G/ LTE`	High correlation
`Touchscreen` is highly imbalanced (90.3%)	Imbalance
`Wi-Fi` is highly imbalanced (94.8%)	Imbalance
`Bluetooth` is highly imbalanced (91.2%)	Imbalance
`GPS` is highly imbalanced (60.0%)	Imbalance
`Number of SIMs` is highly imbalanced (58.4%)	Imbalance
`3G` is highly imbalanced (51.0%)	Imbalance
`Name` has unique values	Unique
`Processor` has 42 (3.1%) zeros	Zeros
`Front camera` has 18 (1.3%) zeros	Zeros
`Operating system` has 1299 (95.6%) zeros	Zeros

Annexure 2: Evaluation Metrics

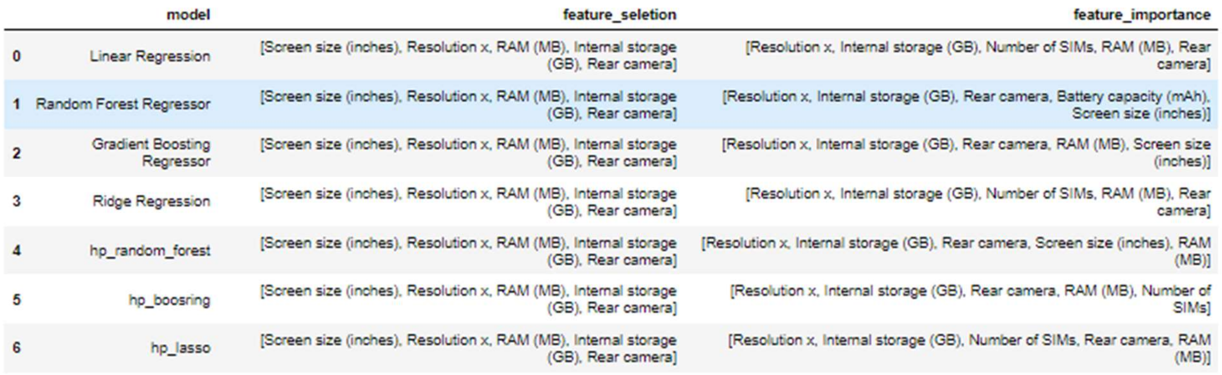




Annexure 3: Hyper Parameter Selection and Feature importance

Parameters selected for tuning and finding optimized model:

Model	Parameters
Random Forest	Number of decision trees and the data points split are considered here to find the best estimations. “n_estimators”,”min_sample_split”, “max_depth”
Gradient Boosting	Similar to the random forest, the learning_rate is consider as additional component to find the best prediction factors. “n_estimators”,”learning_rate”, “max_depth”
Regularization	“alpha” – the regularization constant (strength)



	model	feature_seletion	feature_importance
0	Linear Regression	[Screen size (inches), Resolution x, RAM (MB), Internal storage (GB), Rear camera]	[Resolution x, Internal storage (GB), Number of SIMs, RAM (MB), Rear camera]
1	Random Forest Regressor	[Screen size (inches), Resolution x, RAM (MB), Internal storage (GB), Rear camera]	[Resolution x, Internal storage (GB), Rear camera, Battery capacity (mAh), Screen size (inches)]
2	Gradient Boosting Regressor	[Screen size (inches), Resolution x, RAM (MB), Internal storage (GB), Rear camera]	[Resolution x, Internal storage (GB), Rear camera, RAM (MB), Screen size (inches)]
3	Ridge Regression	[Screen size (inches), Resolution x, RAM (MB), Internal storage (GB), Rear camera]	[Resolution x, Internal storage (GB), Number of SIMs, RAM (MB), Rear camera]
4	hp_random_forest	[Screen size (inches), Resolution x, RAM (MB), Internal storage (GB), Rear camera]	[Resolution x, Internal storage (GB), Rear camera, Screen size (inches), RAM (MB)]
5	hp_boosring	[Screen size (inches), Resolution x, RAM (MB), Internal storage (GB), Rear camera]	[Resolution x, Internal storage (GB), Rear camera, RAM (MB), Number of SIMs]
6	hp_lasso	[Screen size (inches), Resolution x, RAM (MB), Internal storage (GB), Rear camera]	[Resolution x, Internal storage (GB), Number of SIMs, Rear camera, RAM (MB)]

References

Mobile phone Specifications and Prices. (2022, August 14). Kaggle. https://www.kaggle.com/datasets/pratikgarai/mobile-phone-specifications-and-prices

Gadgets 360. (n.d.). *Tech News, Latest Technology, Mobiles, Laptops - Gadgets 360*. https://www.gadgets360.com

Sruthi, L. (2022, January 7). Hyperparameter tuning in linear Regression. - Analytics Vidhya - medium. *Medium*. https://medium.com/analytics-vidhya/hyperparameter-tuning-in-linear-regression-e0e0f1f968a1

Lewinson, E. (2023, July 11). *A comprehensive overview of regression evaluation metrics | NVIDIA Technical blog*. NVIDIA Technical Blog. https://developer.nvidia.com/blog/a-comprehensive-overview-of-regression-evaluation-metrics/

Feature selection. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/feature_selection.html

Daython. (2023, May 14). Mastering the Art of Feature selection: Python techniques for visualizing feature importance. *Medium*. https://medium.com/@daython3/mastering-the-art-of-feature-selection-python-techniques-for-visualizing-feature-importance-cacf406e6b71

Dataset: Link

Working folder: Link

GitHub: https://github.com/NILodio/data-mining