**Intel Unnati Industrial Training – Summer 2023**

SMART MOBILE PHONE PRICE PREDICTION USING MACHINE LEARNING

**Team Cosmo**

**Institution - Nitte Meenakshi Institute of Technology**

Siva Subramaniam J

Department of ECE, NMIT

Amruthavarshini M

Department of ECE, NMIT

Institution Mentor

Dr. Rajesh N

Professor, Dept. of ECE, NMIT

Industry Mentor

Mr. Srivatsa Sinha

# Abstract

Mobile phone is one of the greatest inventions of the twentieth century. The world has seen a huge increase in the use of mobile phones in the last few years. We can use mobile phones to make calls, send messages, play games, take pictures, record videos, access the internet and many more. Smartphones are an important part in the lives of human beings. It is very essential to consider all the factors and the price of the smartphone to know its worth. Our model is built to predict the approximate price considering all possible parameters like processor, storage, camera, battery life, display, connectivity and others. It is very essential to know the features to make a right decision to buy a smart phone worth the required specifications. This project gives a model to solve this problem by using available data related to these features and predict the approximate price of the smart phone. This helps customers know the actual value of the smartphone with the required specifications and also allows owners decide the appropriate price for the smartphone for the features they offer.

# Introduction

The aim of this project is to develop a model that predicts the value of a mobile phone based on its characteristics and to find an ML algorithm that estimates the value most accurately. Using collected data to accurately predict future events is an important aspect of forecasting.One way to do predictive is machine learning. Machine Learning predictions works by using data as input to build and train a prediction model, and the training model is used to predict outcomes from previous data.Supervised machine learning algorithms use data that contains a list of predefined classes, which are attributes that need to be predicted.In our case, price of the mobile phone is the class label. The most widely used purchasing and selling gadgets entity nowadays are smartphones. New cell phones with improved most software and additional features are introduced every day. Mobile phones are bought and sold in huge numbers every day. In light of this, the mobile price class prediction would be a case study for the specific problem type, namely, finding the optimal product.

# Literature Survey

The price prediction of smartphones can be done using a variety of algorithms in Machine Learning. The ability of a computer system to efficiently answer questions is known as Artificial Intelligence. Some of these technologies are classification, regression, supervised and unsupervised learning, decision trees and many more models and algorithms. It is very important to choose the best algorithm to reduce the dataset obtained and choose the best characteristics to predict the approximate price of the required features correctly.

[1] shows how dataset can be partitioned into training data and testing data for good evaluation.The dataset is collected from Kaggle and includes about 20 features to predict the price range.This paper uses Naïve Bayes Classifier, K Nearest Neighbour (KNN) Classifier and Randon Forest Classifier to train the data and make the prediction.This also provides a brief information on the metrics used such as confusion matrix, classification report and accuracy score.This shows that accuracy of model can be improved by preprocessing data methods to get better results.

[2] shows different techniques used to reduce the dimensionality of the data obtained.The reduction methods are Feature Selection, Feature Extraction,Forward and Backward Selection.The classification techniques used are ZeroR algorithm, Naïve Bayes algorithm and J48 Decision tree.This shows that a high accuracy prsdiction can be achieved by mining and analysis of data.

[3] uses a model to segregate prices based on the following features : Brand, Ratings, RAM, ROM, Mobile size, Camera and Battery power.Here the data is extracted by Scrapping of Data from the web for example an E-commerce site. But his data requires a lot of pre-processing and handling the missing values. Random Forest Regressor and Support Vector Regressor are the two models that are used in this blog. Explanatory Data Analysis (EDA) is a method of data analysis used to summarize the data based on the characteristics with visual techniques and pair plots for each feature is obtained. Both the used models provided good results with accuracy greater than 90%.

[4] considers Battery power, Bluetooth, Microprocessor clock speed, Dual SIM support, Camera, Internal Memory, 4G support, Core processors, Weight and Dimensions, Charge duration and Wi-Fi support to predict the price of the smartphone. It discusses data analysis using 4 classification algorithms to predict the price: Random Forest Classifier, Naive Bayes, KNN Classifier and SVM Classifier. It discusses the classification algorithms in detail. Random process classifier is a machine learning technique that combines multiple classifiers to solve complicated problems that uses large number of decision trees. Naive Bayes classifier uses the conditional probabilities that is used to find the likelihood of anything that has already occurred. K Nearest Neighbor is a technique that is used for classification, regression and to fill in missing values and resample datasets.Support Vector Machine is a supervised technique used to find the decision boundary to categorize data points.

[5] This paper is aimed to figure out the problem by using machine learning algorithms like Support Vector Machine, Decision Tree, K Nearest Neighbors and Naïve Bayes to train the mobile phone dataset before making predictions of the price level. It used appropriate algorithms to predict smartphone prices based on accuracy, precision, recall and F1 score. This also uses the preprocessing of data to obtain train and test data.They use SVM, Naïve Bayes, Decision Trees and KNN models to predict the price of the mobile phone.

[6] discusses four different algorithms – Random Forest Classifier, KNN Classifier, Logistic Regression and Decision Tree. Various feature extraction methods are used to choose the best features and shrink the dataset. As a result, the problem was be less numerically complicated. The following steps made up the machine learning prediction model - Data collection, processing, prototype, selection, training, assessment, and hyperparameter tuning.

[7] is a Kaggle webpage from which we have chosen the data for smart mobile phone price prediction. This dataset contains information on the prices of several mobile phones from different brands. It includes details such as the storage capacity, RAM, screen size, camera specifications, battery capacity, and price of each device.

# Data Selection and Exploratory Data Analysis

We have chosen a minimalistic dataset with considerable number of significant features which highly influence the price of the smartphone. The chosen dataset is obtained from a publicly available source on Kaggle website. The dataset has 8 columns and 407 rows. It includes the following:

* Brand: the manufacturer of the phone
* Model: the name of the phone model
* Storage (GB): the amount of storage space (in gigabytes) available on the phone
* RAM (GB): the amount of RAM (in gigabytes) available on the phone
* Screen Size (inches): the size of the phone's display screen in inches
* Camera (MP): the megapixel count of the phone's rear camera(s)
* Battery Capacity (mAh): the capacity of the phone's battery in milliampere hours
* Price ($): the retail price of the phone in US dollars

We can tell that the chosen dataset is good to analyze and apply core concept machine learning lifecycle like feature selection, engineering rather than getting lost in data with huge feature set.This data is further cleaned using the data cleaning process which includes the following :

* Convert Storage to integer
* Convert RAM to integer
* Convert ScreenSize to float
* Convert camera into relevant datatype

The most crucial step in machine learning lifecycle which influences the performance of any model directly is Feature Engineering. In our case, there is a need to convert the camera columns into useful data which is utilized to train dataset like the count of camera whose average value is taken as the CameraAvg which makes a total of 4 camera features available.

Data Visualization is required after preprocessing of data.The obtained data is now visualized and we can draw a few inferences on how the data gets distributed and check for any outliers.In case of presence of outliers we either have to remove the outliers completely or use Huber loss or regression.We can study for both data i.e with outliers and without outliers.

On data visualization, we found a number of outliers which are not helpful to predict the price correctly. In order to overcome the problem of outliers we need to remove the outliers from each brand or use Huber Loss/Huber Regression for the model. Although removing of outliers may lead to overfitting, we use this method to improve accuracy of the model. In our project we will remove the outliers by finding the Inter- Quartile Range, upperbound and lowerbound for each brand of mobiles.

* A heat map is used to analyze the data by obtaining the correlation between the features and helps in training the model.From the heat map we observe that storage has the highest correlation with price among all other features and Battery has a negative correlation with price.This makes it evident that higher priced phones come with a less battery pack.It is also observed that the features do not correlate among themselves which makes it good to go for model training.

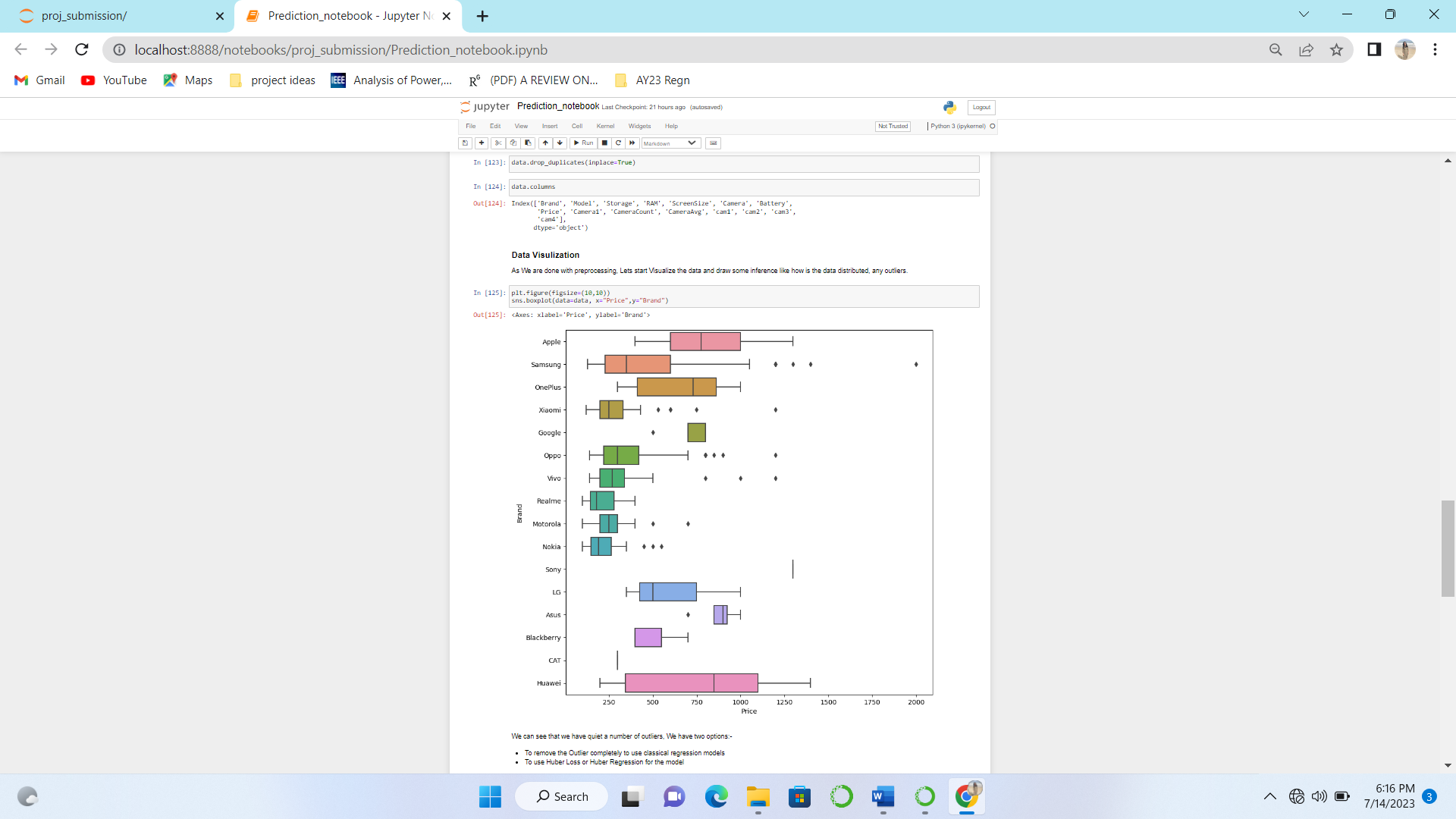
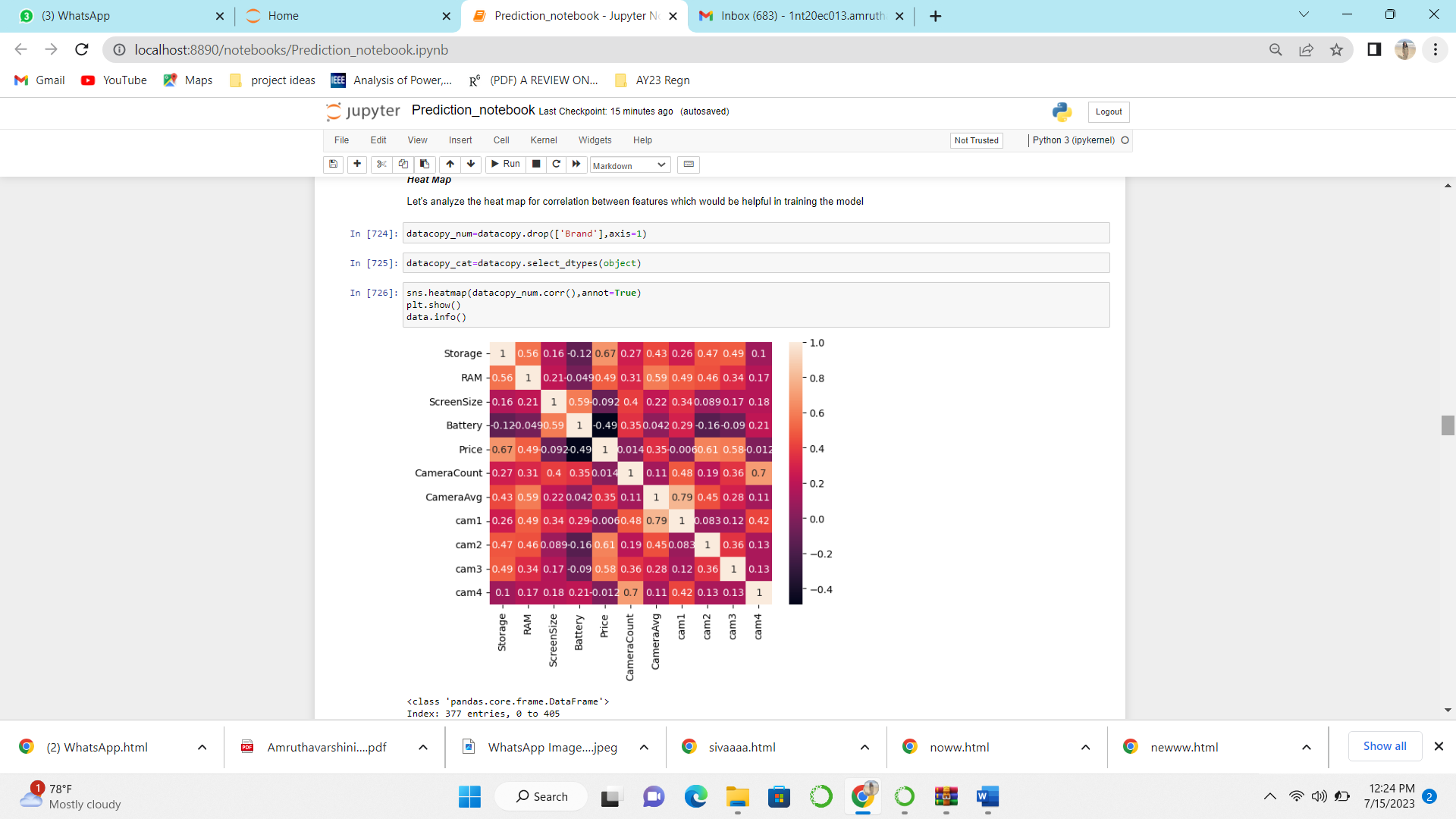
 

Fig. Visualization of Data- Price v/s Brand Fig. Heat Map

As we all know the Brand value helps in predicting the price. Hence we will have to encode it into numbers to help us train the model. We are using the LabelEncoder to encode the Brand value. We have split the data into train and test dataset using the train\_test\_split function in sklearn.

# Metric and Model Selection

We have used three classical metrics used for regression problems.They are :

* Mean Squared Error (MSE) - It is an important loss function for algorithms fit/optimized using the least squares framing of a regression problem where “least squares” refers to minimizing the mean squared error between predictions and expected values.
* Mean Absolute Error (MAE) - The units of the error score match the units of the target value that is being predicted.The changes are linear and hence intuitive.
* Root Mean Squared Error (RMSE) - is an extension of the mean squared error. Importantly, the square root of the error is calculated, which means that the units of the RMSE are same as the original units of the target value that is being predicted.

## Linear Regression Model

We have started using the Classical Linear Regression Model which provides a sloped straight linear representing the relationship between the features. This provides us a reference to predict the value of our target i.e price.The results are analyzed both with outliers and in their absence and it was observed that the Linear Regression model performs well for data without outliers.Therefore, we can infer that Linear Regression model is sensitive to outlier effects and the metric results are not satisfactory.

Without Outliers With outliers

R2\_score is :0.7943246715176291

MAE is :74.97102444927953

MSE is :10460.77592858767

RMSE is :102.27793471021826

R2\_score is :0.7184723338370405

MAE is :114.45593457405009

MSE is :35797.09842689034

RMSE is :189.20121148367508

## Stochastic Gradient Descent Regression

SGD Regression model was supposed to work better than the linear regression but results were observed to be similar.This could be due to the small dataset with about 300 data points to train on. This score could be increased by using cross validation and used best parameters using grid search or random search cross validation. This shows that regression models are affected by the outliers.

Without Outliers With outliers

R2\_score is :0.7945268083318554

MAE is :75.25444683172402

MSE is :10450.495123708728

RMSE is :102.22766320183949

R2\_score is :0.7197179810480463

MAE is :114.96397657096021

MSE is :35638.71059799488

RMSE is :188.78217764925503

## Random Forest Regressor

## It is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. It takes less training time as compared to other algorithms. It predicts output with high accuracy.

Without Outliers With outliers

R2\_score is :0.9170069116580183

MAE is :40.39110513296229

MSE is :4221.080414325757

RMSE is :64.96984234493537

R2\_score is :0.814173598601792

MAE is :82.47048755094808

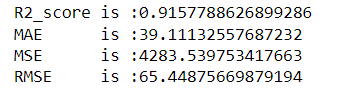
MSE is :23628.391737940303

RMSE is :153.7152944177654

A drastic difference was observed in the results between the dataset with outlier and without outlier. The dataset without outlier has about half the mean absolute error of the original dataset with outliers. Random Forest as mentioned gives about 90% R2 score and a Mean Absolute Error of 40. It is also inferred that the random forest do not overfit to the dataset. The drawback of Random Forest algorithm is that it is not meant for regression as the range of predictions a Random Forest can make is bound by the highest and lowest labels in the training data. This behavior becomes problematic in situations where the training and prediction inputs differ in their range and distribution.

## CatBoost Regressor

Without Outliers



## It is a gradient-boosting ML system that sets itself apart from other Gradient Boosted Decision Trees by offering unique solutions for interpreting data sources that are highly categorical or contain missing data points. CatBoost achieves this by using the split-by-popularity method to create symmetrical decision trees. By grouping features into a single split with only a left and right child, the necessary processing power and time are greatly reduced when compared to trees with children for each individual feature in a set. These features could be categorical or numerical.Similar results were obtained but this does not require encoding of categorical values.

## Light Gradient Boosting Machine(GBM) Regressor

Without outliers

R2\_score is :0.8995773502789502

MAE is :46.54820967097329

MSE is :5107.558814362076

RMSE is :71.46718697669635

Light GBM is a gradient boosting ensemble method that is used by the Train Using AutoML tool and is based on decision trees. Light GBM is optimized for high performance with distributed systems. From the results we can see that there is dip in R2\_score and the MAE, RMSE have increased as compared to RandomForest and CatBoostRegressor. This may be due to the fact that there is only Limited Feature Selection for a given dataset. It will select only the features with highest correlation to the target variable. However, finding the best combination of hyperparameters to maximizing accuracy can be challenging and can be time-consuming.

## XGB (Extreme Gradient Boosting) Regressor

XG Boost (Extreme Gradient Boosting) dominates structured or tabular datasets on classification and regression predictive modeling problems. The two main reasons to use XG Boost are execution speed and model performance. The difference between XGB Regressor and LGBM Regressor is that XGB Regressor uses level-wise(horizontal) growth. LGBM Regressor uses leaf-wise(vertical) growth. The metrics of XGB Regressor are similar to the LGBM Regressor but are better than same. This is due to the better optimization of the XGB Regressor and it works well with sparse input data for tree and linear booster

R2\_score is :0.9025859452430992

MAE is :41.54388711111886

MSE is :4954.5397915552785

RMSE is :70.3884918971509

## Huber Regressor

The Huber Regressor optimizes the squared loss for the samples where |(y - Xw - c) / sigma| < epsilon and the absolute loss for the samples where |(y - Xw - c) / sigma| > epsilon, where the model coefficients w, the intercept c and the scale sigma are parameters to be optimized. The parameter sigma makes sure that if y is scaled up or down by a certain factor, one does not need to rescale epsilon to achieve the same robustness. Result without outliers. Huber Regression also underperforms in the dataset with outliers. Therefore, in any dataset, before training the model it is highly recommended to remove the outliers.

Without Outliers With outliers

R2\_score is :0.8051360786444376

MAE is :67.61827133095615

MSE is :9910.900995795362

RMSE is :99.55350820435893

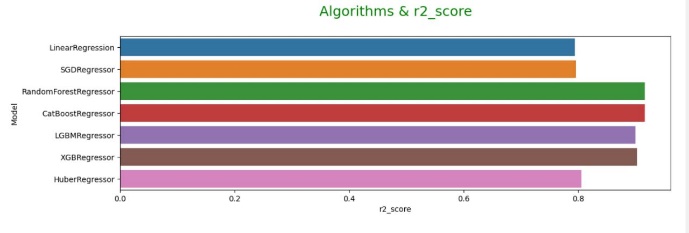
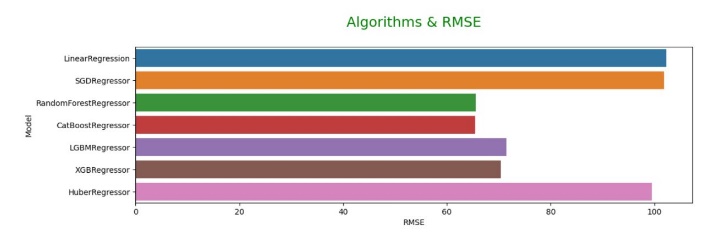
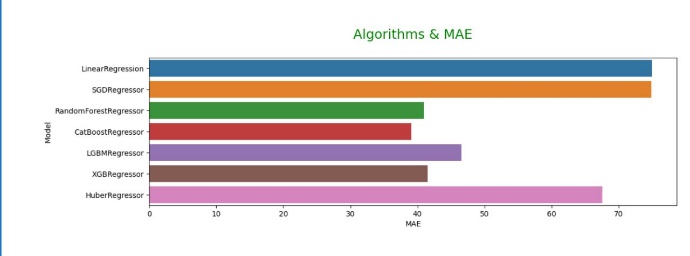
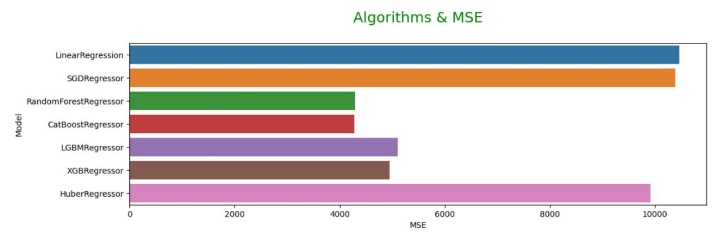
R2\_score is :0.6625565194846426

MAE is :120.50050734107542

MSE is :42906.96417214145

RMSE is :207.1399627598244

# Model Evaluation and Future Works



From the graphs obtained by plotting Metrics against Models, we can observe that Random Forest Regressor performs best among all the other models which is then followed by the LGBM Regressor and XGB Regressor with similar metrics that r2\_score and with a very low magnitude variations in the MSE,MAE and RMSE.

### Additional Ideas for future work

* Increase the model performance by using Cross Validation
* Using any of the methods such as Grid Search CV or Random Search CV to estimate best hyperparameters for the model which could improve the performances of the model.

# Conclusion

From the graphs and the metrics we can conclude that RandomForestRegressor and CatBoostRegressor performs best among all the other models.

It is followed by the LGBMRegressor and XGBRegressor with similar metrics that r2\_score and with a very low magnitude variations in the MSE,MAE and RMSE.

It might be surprising to see a Categorical algorithm performing better than the models which are specially designed for regression for a regression problem

One reason is fact that the features used such RAM, Memory and Storage are indirectly an categorical data as there are only discrete values for them. That helps the CatBoostRegression and RandomForestRegression Decision Trees to make quality decision.

Also It is fact that the dataset is sparse for the specialized Regression Models to show a considerably higher performance than CatBoostRegressor or RandomForestRegressor

Currently we can draw conclusion that RandomForestRegressor and CatBoostRegressor performs better compared to other Regression Model

When the size of dataset is increased we may see the specilized Regression Models outperforming the forementioned models. This maybe taken as future works on the project.

# References

[1] Kiran, A. Varun, and R. Jebakumar. "Prediction of Mobile Phone Price Class using Supervised Machine Learning Techniques." *International Journal of Innovative Science and Research Technology* 7 (2022): 248-251. ----20 features

[2] Arora, Pritish, Sudhanshu Srivastava, and Bindu Garg. "Mobile price prediction using weka 1." *vol* 5 (2020): 330-333. ----J48

[3] Nivitus, “medium,” 22 July 2020. [Online]. Available:https://medium.com/@Nivitus./mobile-price-prediction-using-machine-learning-fa9cab6fb242 ----EDA

[4] P. Majumder, "Learn Mobile Price Prediction Through Four Classification Algorithms," Analytics Vidhya, March 2022. [Online]. Available: https://www.analyticsvidhya.com/blog/2022/02/learn-mobile-price-prediction-through-four-classification-algorithms/

[5] N. Hu, "Classification of Mobile Phone Price Dataset Using," in 3rd International Conference on Pattern Recognition and Machine Learning, 2022.

[6] Rao, M. Prashanth, et al. "Price to Performance Ratio of Smart Phones using Machine Learning." *2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*. IEEE, 2023.

[7] K.RATTANAPORN, "Mobile Phone Price," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/rkiattisak/mobile-phone-price.