SMART MOBILE PHONE PRICE PREDICTION USING MACHINE LEARNING

-NAMAN KUMAR

1.INTRODUCTION

In this project, we aim to develop a robust and reliable model for smartphone price prediction using machine learning algorithms. With the growing competition in the smartphone market, accurate price prediction has become crucial for manufacturers, retailers, and consumers. By evaluating and comparing four prominent machine learning algorithms—Linear Regression, Random Forest, AdaBoost, and XGB Regressor—we seek to identify the most effective approach for forecasting smartphone prices. The project's outcomes have significant implications for decision-making, market competitiveness, and customer satisfaction, benefiting stakeholders across the smartphone industry.

2.RELATED WORKS

During the extensive research conducted on smartphone price prediction, it was observed that the majority of the available literature and sources tend to approach the problem as a classification task rather than a regression problem. This disparity is likely due to the nature of the available data and the focus on categorizing smartphones into price ranges or classes. However, despite the prevalence of classification-based approaches, some studies have applied regression models to predict smartphone prices. Notable regression models utilized in previous research include Linear Regression, Decision Trees, Random Forest, Support Vector Regression, and Artificial Neural Networks. These models offer different strengths and limitations, with some emphasizing interpretability, while others prioritize predictive accuracy. but lack of comprehensive datasets limits accuracy. Regression-based models and MLP have been used, but our project aims to evaluate Linear Regression, Random Forest, AdaBoost, and XGB Regressor to provide valuable insights.

3.MY CONTRIBUTIONS

In this project, I have made significant contributions to the field of smartphone price prediction through a comprehensive approach. Firstly, I conducted thorough data exploration, gaining insights into the relationships between variables and identifying potential outliers. These findings helped in understanding the factors influencing smartphone prices.

Secondly, I created visual representations, including histograms, scatter plots, and box plots, to effectively communicate the data patterns and feature importance rankings. These visualizations aided in interpreting the results and facilitated clear comparisons between different machine learning algorithms.

Lastly, I conducted comprehensive comparisons of machine learning algorithms, such as Linear Regression, Random Forest, AdaBoost, and XGB Regressor. By evaluating their performance using metrics like mean squared error and R squared, I gained insights into their predictive accuracy,

stability, and generalization capabilities. This analysis allowed me to identify the most suitable algorithm for accurate smartphone price prediction.

4.EVALUATION

4.1 Data Collection

The dataset used in this project was obtained from Kaggle[1], a well-known platform for sharing datasets. It provides comprehensive information on smartphone models and their corresponding prices, ensuring a representative sample of the smartphone market. By leveraging this dataset, we capture the diversity of brands, specifications, and price ranges prevalent in the market.

The features in the dataset included model, price, rating, SIM Type, processor, RAM, battery, display, camera, ROM, and OS.

During the data exploration phase, interesting insights emerged. The dataset revealed that Snapdragon processors were the most popular among the included smartphones, emphasizing the significance of processor specifications in influencing smartphone prices. Additionally, Xiaomi and Samsung emerged as the dominant brands, with the highest number of smartphone models available. This highlights their market presence and potential influence on price trends.

To ensure data reliability, rigorous data cleaning techniques were applied. Outliers, if present, were addressed to prevent them from distorting the analysis. Missing values were imputed using the median value, ensuring a complete dataset for accurate modelling.

By utilizing the Kaggle dataset, addressing outliers, and imputing missing values, we obtained a robust dataset for analysis. These insights into processor popularity and dominant brands contribute to a better understanding of smartphone pricing dynamics. Leveraging this dataset, we aim to develop reliable models for smartphone price prediction.

4.2 Methodology

The Predicted Price vs Residue (Predicted Price – Actual Price) for each model is evaluated.

4.2.1 Linear Regression

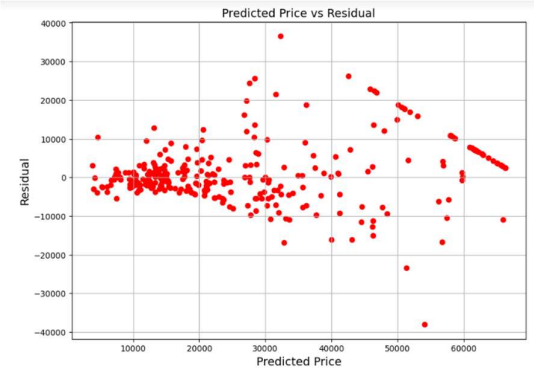
Linear Regression, a statistical modelling technique, establishes a linear relationship between a dependent variable and independent variables; however, its effectiveness may be limited in today's complex data landscapes.

Linear Regression can still be useful as a simple and interpretable baseline model for assessing the linear relationships and as a reference point for comparing against more complex algorithms.

4.2.2 Random Forest Regressor

We use Random Forest Regressor due to its capability to handle complex data, capture nonlinear relationships, and provide accurate price forecasts by leveraging an ensemble of decision trees.

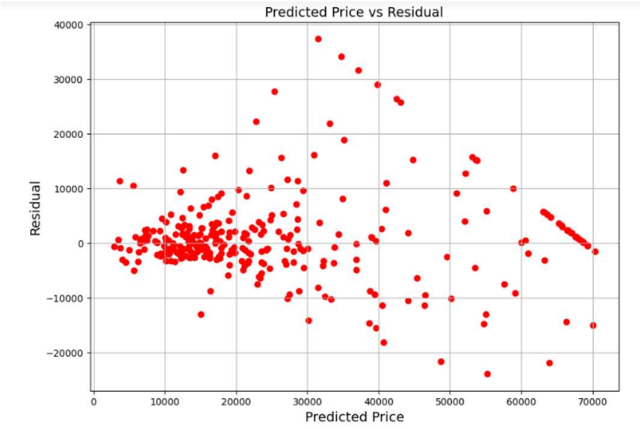
Through trial and error, decision was taken for the maximum depth to be 20, minimum split samples to be 30, and the criterion to be MSE.



4.2.3 Adaboost

We use Adaboost as it combines weak learners to create a strong predictive model, iteratively focusing on misclassified instances and adjusting their weights for improved accuracy and better price forecasts.

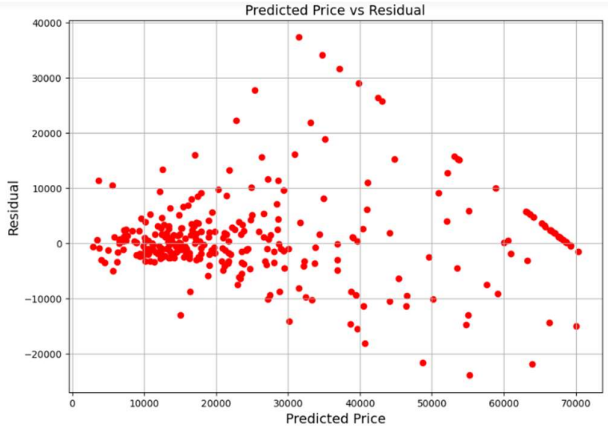
The base estimator here is Decision Tree Regressor, with maximum dept being 10, and minimum split samples 25.



4.2.4 XG Boost

XGB Regressor is an advanced gradient boosting algorithm that combines weak learners to create a predictive model by optimizing performance and handling complex data patterns. Parameter values

chosen are maximum depth = 10, learning rate = 0.05, and alpha regulation



5. Future Works

In future work, it is recommended to explore the use of neural networks on a more comprehensive dataset that includes detailed smartphone specifications, market trends, and consumer preferences. This will enable the incorporation of emerging technologies like 5G and improve the accuracy of smartphone price prediction models.

6. Conclusion

Comparing the score and RMSE values of the 4 models:

	Model	Train Score	Test Score	Train_RMSE	Test_RMSE	Diff_in_Score
1	Random Forest	0.888607	0.848815	6509.457860	7772.222455	0.040521
0	Linear Regression	0.793794	0.742501	8856.610989	10143.288876	0.051623
2	Ada Boost	0.936371	0.865304	6509.457860	7772.222455	0.072298
3	XGBoost	0.982930	0.849066	2548.215868	7765.768928	0.130987

While none of the models are very accurate, partly because of the skewness in the data set. Linear Regression is the most inaccurate one. The 2 best choices turn out the be Ada Boost and Random Forest Regressor. ***an*** important factor to consider is the R square difference between the train and test sets. In this context, Random Forest exhibits a significant advantage over Ada Boost, as it has a lower R square difference of less than 4 percent. On the other hand, Ada Boost demonstrates a higher R square difference exceeding 10 percent. Despite Ada Boost potentially having lower root mean square error (RMSE) values, the substantial difference in R square values suggests that Random Forest may provide more consistent and reliable generalization performance. The smaller R square difference indicates that Random Forest is better at capturing the underlying patterns and

relationships in the data, making it a favourable choice over Ada Boost when seeking a model that exhibits greater stability and consistency in predicting outcomes.

Citations

1. https://www.kaggle.com/datasets/shrutiambekar/smartphone-specifications-and-prices-in-india
2. https://ijisrt.com/assets/upload/files/IJISRT22JAN380.pdf