International Journal of Innovative Research in Engineering
Volume 4, Issue 2 (March-April 2023), PP: 603-607.
www.theijire.com ISSN No: 2582-8746

Smartphone Price Prediction Using Machine Learning Techniques

Richard Honey
Department of Economics, Christ (Deemed to be University), Bangalore, Delhi NCR Campus, India.

How to cite this paper:
Richard Honey, "Smartphone Price Prediction Using Machine Learning Techniques", IJIRE-V4I02-603-607.

https://www.doi.org/10.59256/ijire.2023040233

Abstract: *This research aimed to predict smart phone prices using two supervised machine learning algorithms: Decision Tree and Random Forest Regression. Data was collected from the Indian e-Commerce website Flip kart using Python libraries such as Beautiful Soup and Selenium, and was cleaned and pre-processed for analysis. The results showed that the Decision Tree algorithm had an $R^2$of 89.3%. The Random Forest classifier showed the R^2 value with an accuracy score of 82.8%. The study offers a method for accurately predicting smart phone prices that could be useful to determine the cost of their products and ultimately benefit the entire smart phone market.*

Key Word: *Smartphone, Price Prediction, Machine Learning, Decision Tree, Random Forest Regression.*

I.INTRODUCTION

The smartphone industry has grown exponentially over the past decade, with new models being introduced every year. With the increasing competition, it has become crucial for smart phone manufacturers to determine the cost of their products accurately. The goal of this study is to develop an accurate model that will aid smart phone manufacturers in determining the cost of smartphones. The study aims to predict smart phone prices using two supervised machine learning algorithms: Decision Tree Regression, and Random Forest Regression.

Smartphone price prediction has been a topic of interest in the research community, with several studies focusing on the development of accurate models to predict smart phone prices. Many of these studies use sample datasets from Kaggle data science competitions to predict smart phone prices (Subhiksha, Thota& Sangeetha, 2019). Another group of researchers makes use of data collected from online stores, such as GSMArena, to predict smart phone prices (Azim & Khan, 2018; Salmasi et al. 2021). Several studies often fail to take into account important factors for price prediction, such as 5G connection and advanced display types.

This study addresses the gap in the existing literature by considering these important factors in the analysis. The study uses data collected from the Indian e-commerce website Flip kart and applies various data mining techniques, such as Beautiful Soup and Selenium, to extract relevant information. The data is then cleansed and transformed, and features are classified. The study then applies decision tree and Random Forest classification to the dataset, using the scikit-learn and Grid Search CV packages in Python.

The use of multiple machine learning algorithms in this study provides a more comprehensive analysis of the data, allowing for a more accurate prediction of smart phone prices. The results of the study will be useful not only for smart phone manufacturers but also for the entire smart phone market.

II.REVIEW OF LITERATURE

Although there is a long history of price prediction in diverse fields, there are few studies focused on product price prediction, particularly smart phone price prediction, which is the major goal of this study. Prediction of mobile price classes (Asim& Khan, 2018) involves dividing prices into four categories, in this study. Data was gathered from GSM Arena, and after feature transformation, Naive Bayes and Decision Tree Classifiers were used to classify the data.

Multiple linear regression is also used in another study (Surjuse et al. 2022) to estimate laptop prices, and this approach has an accuracy rate of 81%. Data for the laptop price prediction was feature engineered and taken from Kaggle. The authors of the research Quader&Gani (2017) have acquired historical data from several sources, including IMDB and Rotten Tomatoes. The authors used SVM and neural networks to implement after processing the data to determine the prediction's accuracy. The neural network provided the most precise predictions for the data set that the team picked.

In contrast to earlier models, the study (Salmasi et al. 2021) employed a multimodel strategy with five deep learning models to estimate smart phone prices using data collected from GSMArena. The deep learning technique utilized is called Convolutional Neural Network (CNN). The experimental outcomes demonstrate an F1-score of 88.3%, which indicates that multimodal learning produces more precise predictions. Another study (Xu et al. 2019) predicted water price using three Random Forest Regression model. Cross-validation is applied to evaluate the model performance in the dataset from 1987 to

2009. Result shown a good prediction with higher degree of accuracy without leading to over fitting.

A housing prediction model (Sawant et al. 2018) for housing sector in India is developed on the basis of Decision Tree and Random Forest Regression. Random Forest provides higher accuracy than Decision Tree algorithm and Gini importance is estimated. An effective and precise method for forecasting the cost of a smart phone has been established using the literature.

The paper by Pudaruth (2006) predicted the price of used cars using supervised machine learning algorithms like multiple linear regression, random forest regression, naïve bayes, and decision trees. However, the study used less number of observations.

A dataset used to predict the price of mobile phones that includes 21 parameters gathered from Kaggle. Subhiksha, Thota, & Sangeetha (2020) used Support Vector Machine, Logistic Regression, and Random Forest Classifier as their classification models. The most accurate technique was the support vector machine (81%) and Random Forest Classifier gave an accuracy of 0.77. The publication (Tziridis et al. 2017) also addresses the issue of predicting flight costs. For training ML regression models including Random Forest Regression Tree, a dataset of 1814 Aegean Airlines flights for a certain foreign destination is created. This regression issue can be handled by ML models with almost 88% accuracy.

The study (Pushpa et al. 2017) predicts the class students' results using a classification machine learning method. The Support Vector Machine, Naive Bayes Classifier, Random Forest Classifier, and Gradient Boosting Algorithm are just a few examples of the supervised learning algorithms employed. Accuracy of model in different number of trees analysed in the paper.

Additionally, the authors (Chandrashekhara et al. 2019) used the dataset gathered from e-commerce websites to forecast the price of smart phones following feature transformation. Machine learning methods include linear regression, back propagation neural networks, and support vector regression. The major goal of putting up these three models is to determine which algorithm produces results for price prediction that are more precise. SVR provided reliable findings with 86%.

The majority of these studies used tree based machine learning algorithms. However, only a few studies have considered the impact of factors such as 5G connection and advanced display types on smart phone prices and used datasets available in e-commercial websites. This study addresses these limitations by using multiple machine learning algorithms and considering important factors in the analysis.

III. METHODOLOGY

Dataset: Listed below are the various features in the table.

- Operating System: Operating systems (windows, android and iOS) used in Smartphone and dummy variables used (has 3 columns).
- Number of Sim Slots
- Ratings: Rating to the phone out of 5
- Number of Ratings
- Reviews: Number of reviews
- RAM
- Storage
- Expandable Storage: Expandable up to how many GB.
- Expandable or Not: Storage is expandable or not
- Warranty: Warranty period of product in days.
- Price
- Front Camera1: First front camera
- Front Camera2: Second front camera
- Number of Front Cameras: How many front cameras for the phone.
- Display_Size: Size of the display.
- Display Type: Different types of displays as columns and dummy variable used (has 20 columns).
- Battery_(mAh): Battery power of the product.
- Rear Camera1: First back camera
- Rear Camera2: Second back camera
- Rear Camera3: Third back camera
- Rear Camera4: Fourth back camera
- Number of Rear Cameras
- 5G: 5G is supported or not

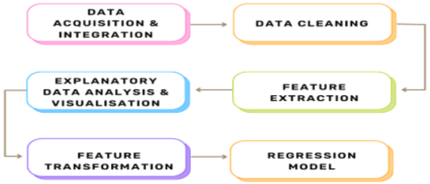
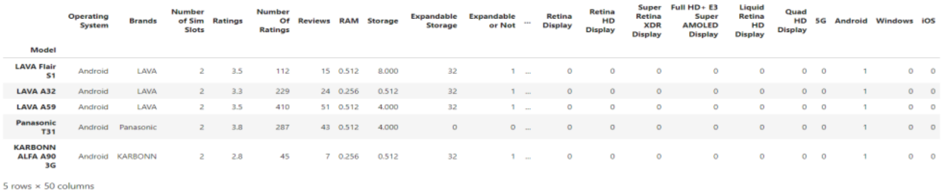

DATA ACQUISITION & INTEGRATION
DATA CLEANING
EXPLANATORY DATA ANALYSIS & VISUALISATION
FEATURE EXTRACTION
FEATURE TRANSFORMATION
REGRESSION MODEL

Figure. 1: Research Workflow.

Work flow: The study uses data collected from the Indian e-commerce website Flip kart, which was mined using the Beautiful

Soup and Selenium packages in Python. The data was then cleansed and transformed, and features were classified. Data analysis & visualization using Python libraries including pandas, Matplotlib and Seaborn is conducted. Feature transformation of every features performed using z-score transformation. The data is divided into training and test data. Next process involved applying Decision Tree Regression and Random Forest Regression to our dataset. Scikit learn and Grid Search CV packages of python used to fit and predict the Random Forest Regression and Decision Tree Regression model.

IV.RESULT

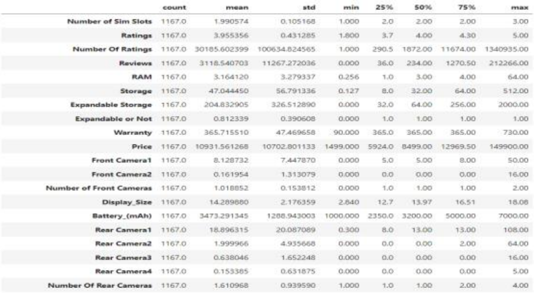


	Operating System	Brands	Number of Sim Slots	Ratings	Number Of Ratings	Reviews	RAM	Storage	Expandable Storage	Expandable or Not	...	Retina Display	Retina HD Display	Super Retina XDR Display	Full HD+ E3 Super AMOLED Display	Liquid Retina HD Display	Quad HD Display	5G	Android	Windows	iOS
Model																					
LAVA Flair S1	Android	LAVA	2	3.5	112	15	0.512	8.000	32	1	...	0	0	0	0	0	0	0	1	0	0
LAVA A32	Android	LAVA	2	3.3	229	24	0.256	0.512	32	1	...	0	0	0	0	0	0	0	1	0	0
LAVA A59	Android	LAVA	2	3.5	410	51	0.512	4.000	32	1	...	0	0	0	0	0	0	0	1	0	0
Panasonic T31	Android	Panasonic	2	3.8	287	43	0.512	4.000	0	0	...	0	0	0	0	0	0	0	1	0	0
KARBONN ALFA A90 3G	Android	KARBONN	2	2.8	45	7	0.256	0.512	32	1	...	0	0	0	0	0	0	0	1	0	0

5 rows × 50 columns

Figure. 2: The data.

.



	count	mean	std	min	25%	50%	75%	max
Number of Sim Slots	1167.0	1.990574	0.105168	1.000	2.0	2.00	2.00	3.00
Ratings	1167.0	3.955356	0.431285	1.800	3.7	4.00	4.30	5.00
Number Of Ratings	1167.0	30185.602399	100634.824565	1.000	290.5	1872.00	11674.00	1340935.00
Reviews	1167.0	3118.540703	11267.272036	0.000	36.0	234.00	1270.50	212266.00
RAM	1167.0	3.164120	3.279337	0.256	1.0	3.00	4.00	64.00
Storage	1167.0	47.044450	56.791336	0.127	8.0	32.00	64.00	512.00
Expandable Storage	1167.0	204.832905	326.512890	0.000	32.0	64.00	256.00	2000.00
Expandable or Not	1167.0	0.812339	0.390608	0.000	1.0	1.00	1.00	1.00
Warranty	1167.0	365.715510	47.469658	90.000	365.0	365.00	365.00	730.00
Price	1167.0	10931.561268	10702.801133	1499.000	5924.0	8499.00	12969.50	149900.00
Front Camera1	1167.0	8.128732	7.447870	0.000	5.0	5.00	8.00	50.00
Front Camera2	1167.0	0.161954	1.313079	0.000	0.0	0.00	0.00	16.00
Number of Front Cameras	1167.0	1.018852	0.153812	0.000	1.0	1.00	1.00	2.00
Display_Size	1167.0	14.289880	2.176359	2.840	12.7	13.97	16.51	18.08
Battery_(mAh)	1167.0	3473.291345	1288.943003	1000.000	2350.0	3200.00	5000.00	7000.00
Rear Camera1	1167.0	18.896315	20.087089	0.300	8.0	13.00	13.00	108.00
Rear Camera2	1167.0	1.999966	4.935668	0.000	0.0	0.00	2.00	64.00
Rear Camera3	1167.0	0.638046	1.652248	0.000	0.0	0.00	0.00	16.00
Rear Camera4	1167.0	0.153385	0.631875	0.000	0.0	0.00	0.00	5.00
Number Of Rear Cameras	1167.0	1.610968	0.939590	1.000	1.0	1.00	2.00	4.00

Figure. 3: Descriptive statistics of continuous variables

.

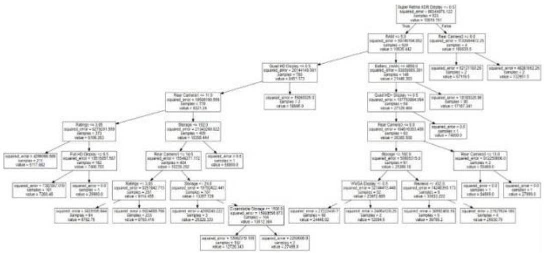


Figure. 4: Decision Tree diagram of smartphone price prediction

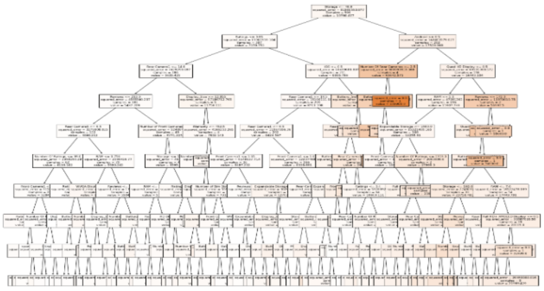


Figure. 5: First Decision Tree of Random Forest Regression model.

The R^2value, which represents the goodness of fit of the model, lies between 0 and 1. A higher R^2value indicates a better fit of the model to the data. The correlation value, which measures the strength of the relationship between variables, also contributes to the evaluation of the model's performance. In general, values closer to 1 are considered more efficient, indicating a better model fit.

Comparing the Random Forest Regression (RFR) and Decision Tree Regression (DTR) models, it was observed that the TheR^2 value, which represents the goodness of fit of the model, lies between 0 and 1. A higher R^2 value indicates a better fit of the model to the data. The correlation value, which measures the strength of the relationship between variables, also contributes to the evaluation of the model's performance. In general, values closer to 1 are considered more efficient, indicating a better model fit.

Comparing the Random Forest Regression (RFR) and Decision Tree Regression (DTR) models, it was observed that the R^2 values for RFR and DTR were 0.89 and 0.82, respectively. Since the R^2 value for RFR was closer to 1, it was concluded that RFR was more effective than DTR in predicting prices of mobile phones.

To visually compare the predicted and actual prices of mobile phones, a graph was plotted for each phone using both models. In the graph, the actual prices were represented in blue, while the predicted prices were represented in green. The overlapping points between the blue and green lines indicate that the actual and predicted prices were very close, while an increasing gap between the lines suggests a larger difference between the predicted and actual prices. The graph for RFR showed a higher degree of overlap compared to DTR.

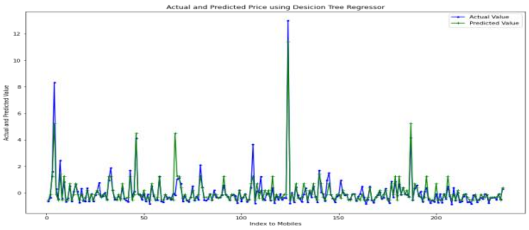


Figure. 6: Actual price and predicted price vs index to mobiles of Decision Tree Regression.

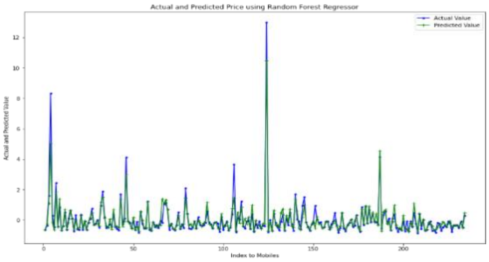


Figure. 7: Actual price and predicted price vs index to mobiles of Random Forest Regression.

To further evaluate the performance of RFR and DTR, error values were calculated and compared in the table below. It was found that RFR had lower error values, indicating more accurate predictions compared to DTR.

Table1: Evaluation metrics of Decision Tree Regression and Random Forest Regression.

Metrics Algorithm	R^2	MEA	MSE	RMSE	MAPE
Decision Tree Regression	0.828	0.265	0.267	0.516	1.25
Random Forest Regression	0.892	0.206	0.167	0.409	0.897

In addition, the Gini importance, which measures the feature importance in Random Forest models, was analyzed for both RFR and DTR. Out of the 44 features, 40 independent variables in RFR scored in different ranges of Gini importance, while only 15 features in DTR scored in different ranges. Therefore, it was concluded that RFR had a better performance in terms of Gini importance.

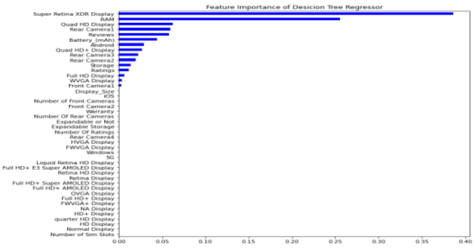


Figure. 8: Gini importance plotted as vertical bar chart (Decision Tree Regression).

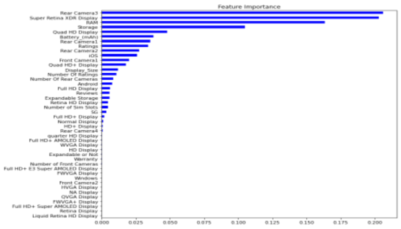


Figure. 9: Gini importance plotted as vertical bar chart (Random Forest Regression).

In summary, the results of the research suggest that RFR is more effective than DTR in predicting prices of mobile phones, as indicated by the higher R^2 value, lower error values, and better performance in Gini importance.

V.CONCLUSION

After conducting extensive research on mobile phone price prediction using machine learning algorithms, it can be concluded that Random Forest Regression is a better approach than Decision Tree Regression. The analysis ofR^2values and correlation coefficients indicated that Random Forest Regression provides more efficient results, as the R^2value was closer to 1 compared to Decision Tree Regression. Moreover, the graphs of actual and predicted values for each mobile phone showed that Random Forest Regression lines overlapped more, indicating a higher degree of accuracy in predicting mobile phone prices. This was further supported by the error values, where Random Forest Regression exhibited lower errors compared to Decision Tree Regression. The higher number of variables considered by Random Forest Regression led to better performance in Gini Importance. These findings suggest that Random Forest Regression is a more accurate and robust approach for predicting mobile phone prices.

In conclusion, this research demonstrated that machine learning algorithms can be effective in predicting mobile phone prices. The outcomes of this study can have important implications for the mobile phone industry, particularly for pricing decisions and market positioning. The findings of this study can be useful for various stakeholders in the mobile phone industry, such as manufacturers, retailers, and consumers. Manufacturers can use the prediction model to estimate the prices of new mobile phone models based on their specifications. Retailers can use the model to set competitive prices for their products, and consumers can use it to make informed decisions about purchasing a mobile phone. Moreover, the research contributes to the literature on machine learning applications in the field of pricing by providing insights into the effectiveness of different algorithms for predicting mobile phone prices. Future research can explore the applicability of Random Forest Regression to other industries and domains, and compare it with other prediction methods to identify the most effective approach for various contexts.

References

[1]. *Asim, M., & Khan, Z. (2018, March). Mobile Price Class prediction using Machine Learning Techniques. International Journal of Computer Applications (0975 – 8887) Volume 179 – No.29. https://doi.org/10.5120/ijca2018916555*

[2]. *Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324*

[3]. *Cavallo, D., Stocchero, M., Gorrasi, G., Logrieco, A., &Attolico, G. (2017). Contactless and non-destructive chlorophyll content prediction by random forest regression: A case study on fresh-cut rocket leaves. Computers and Electronics in Agriculture, 140, 303–310. https://doi.org/10.1016/j.compag.2017.06.012*

[4]. *Chandrashekhara, K. T., M, T., Babu, C. N. G., &Manjunath, T. N. (2019). Smartphone Price Prediction in Retail Industry Using Machine Learning Techniques. Emerging Research in Electronics, Computer Science and Technology. Lecture Notes in Electrical Engineering, vol 545. https://doi.org/10.1007/978-981-13-5802-9_34*

[5]. *Hjerpe, A. (2016). Computing Random Forests Variable Importance Measures (VIM) on Mixed Continuous and Categorical Data. KTH Royal Institute of Technology School of Computer Science and Communication. https://www.diva-portal.org/smash/get/diva2:921542/FULLTEXT01.pdf*

[6]. *Liaw, A., & Wiener, M. (2022). Classification and Regression by Random Forest. R News, 2. https://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf*

[7]. *Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., &Hamprecht, F. A. (2009). A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics, 10(1). https://doi.org/10.1186/1471-2105-10-213*

[8]. *Pekel, E. (2020). Estimation of soil moisture using decision tree regression. Theoretical and Applied Climatology, 139(3–4), 1111–1119. https://doi.org/10.1007/s00704-019-03048-8*

[9]. *Pudaruth, S. (n.d.). Predicting the Price of Used Cars using Machine Learning Techniques. International Journal of Information & Computation Technology, ISSN 0974-2239 Volume 4, Number 7, 753–764.*

[10]. *Pushpa, S. K., Manjunath, T. N., Mrunal, T. V., Singh, A., &Suhas, C. (2017). Class result prediction using machine learning. Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/smarttechcon.2017.8358559*

[11]. *Quader, N., &Gani, M. O. (2017). A machine learning approach to predict movie box-office success. Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/iccitechn.2017.8281839*

[12]. *Sawant, R., Jangid, Y. K., Tiwari, T. K., Jain, S., & Gupta, A. (2018a). Comprehensive Analysis of Housing Price Prediction in Pune Using Multi-Featured Random Forest Approach. International Conference on Computing Communication Control and Automation. https://doi.org/10.1109/iccubea.2018.8697402*

[13]. *Subhiksha, S., Thota, S., & Sangeetha, J. (2019). Prediction of Phone Prices Using Machine Learning Techniques. Advances in Intelligent Systems and Computing, Volume 1079. https://doi.org/10.1007/978-981-15-1097-7_65*

[14]. *Surjuse, V., Lohakare, S., Barapatre, A., &Chapke, A. (2022, January). Laptop Price Prediction using Machine Learning. International Journal of Computer Science and Mobile Computing, Vol. 11, Issue. 1, January 2022, Pg.164 – 168. https://doi.org/10.47760/ijcsmc.2022.v11i01.021*

[15]. *Tziridis, K., Kalampokas, T., Papakostas, G. A., &Diamantaras, K. I. (2017). Airfare prices prediction using machine learning techniques. Institute of Electrical and Electronics Engineers. https://doi.org/10.23919/eusipco.2017.8081365*

[16]. *Xu, Z., Lian, J., Bin, L., Hua, K., Xu, K., & Chan, H. Y. (2019). Water Price Prediction for Increasing Market Efficiency Using Random Forest Regression: A Case Study in the Western United States. Water, 11(2), 228. https://doi.org/10.3390/w11020228*

[17]. *Zehtab-Salmasi, A., Feizi-Derakhshi, A., Nikzad-Khasmakhi, N., Asgari-Chenaghlu, M., &Nabipour, S. (2020). Multimodal Price Prediction. Annals of Data Science. https://doi.org/10.1007/s40745-021-00326-z*

[18]. *Zhong, S., Xie, X., & Lin, L. (2015). Two-layer random forests model for case reuse in case-based reasoning. Expert Systems With Applications, 42(24), 9412–9425. https://doi.org/10.1016/j.eswa.2015.08.005*