Smart Mobile Price Prediction Using Machine Learning

Team Name *– The Stinsons* ***Date of Submission –*** *14 July, 2023*

Vellenki Sahith Reddy
Department of Emerging Technologies
Malla Reddy College of Engineering and Technology
Hyderabad, Telangana
20n31a6958@mrcet.ac.in

Vanguri Ghan Shyam Ruthik Rajan
Department of Emerging Technologies
Malla Reddy College of Engineering and Technology
Hyderabad, Telangana
20n31a6955@mrcet.ac.in

***Abstract*— This project focuses on predicting mobile phone prices using machine learning techniques. Data is obtained by web scraping from "www.91mobiles.com" and includes features such as brand, RAM, internal memory, processor, camera specifications, and battery capacity. Preprocessing steps handle missing values and convert categorical variables to numerical representations. Random Forest Regression and Support Vector Regression algorithms are trained and evaluated to predict prices. Performance evaluation metrics and data visualization techniques, such as line plots and heatmaps, provide insights into price variations. The results demonstrate accurate price prediction using the Random Forest Regression model. This project aids in understanding mobile phone pricing and supports decision-making for consumers, manufacturers, and retailers.**

Keywords*— *Mobile phone price prediction, machine learning, web scraping, feature selection, random forest regression, support vector regression, data visualization.

I. Introduction

Mobile phones have become an integral part of our modern society, serving as essential communication devices and multi-functional tools. As the mobile phone market continues to expand with a plethora of options, accurately predicting the price of a mobile phone is crucial for both consumers and businesses. In this project, we aim to develop a machine learning model for mobile phone price prediction, leveraging a dataset obtained through web scraping from the "www.91mobiles.com" website.

The dataset encompasses a comprehensive set of features, including brand, RAM, internal memory, processor, camera specifications, battery capacity, and more. By analyzing these features, we can identify the key factors that influence mobile phone prices and build a predictive model to estimate their values. Through data preprocessing techniques, we handle missing values, convert categorical variables into numerical representations, and ensure the data is in a suitable format for modelling.

To achieve accurate price prediction, we employ two machine learning algorithms: Random Forest Regression and Support Vector Regression. These models are trained on the dataset, and their performance is evaluated using metrics such as mean squared error. By leveraging these algorithms, we can capture complex relationships between features and prices, allowing us to make reliable predictions.

Data visualization plays a vital role in understanding the patterns and trends within the dataset. We employ various visualization techniques, including line plots and heatmaps, to explore the relationships between different features and price variations.

The outcomes of this project have practical implications for both consumers and industry stakeholders. Consumers can utilize the predictive model to make informed purchasing decisions, considering factors such as brand, specifications, and pricing. Manufacturers and retailers can leverage the insights gained to optimize their pricing strategies, aligning them with market trends and consumer preferences.

In conclusion, this project combines web scraping, machine learning, and data visualization to develop a robust mobile phone price prediction model. By accurately estimating mobile phone prices, we aim to empower consumers and industry players with valuable insights, enhancing decision-making processes in the dynamic mobile phone market.

II. Previous Work

Using previous data to predict price of available and new launching product is an interesting research background for machine learning researchers. Sameerchand-Pudaruth [1] predict the prices of second-hand cars in Mauritius. He implemented many techniques like Multiple linear regression, k-nearest neighbours (KNN), Decision Tree, and Naïve Bayes to predict the prices. Sameerchand-Pudaruth got Comparable results from all these techniques. During research it was found that most popular algorithms i.e Decision Tree and Naïve Bayes are unable to handle, classify and predict Numerical values. Number of instances for his research was only 97(47 Toyota+38 Nissan+12 Honda). Due to a smaller number of instances used, very poor prediction accuracies were recorded [1].

Shonda Kuiper [2] has also worked in the same field. Kuiper used multivariate regression model to predict price of 2005 General Motor cars. He collected the data from available online source www.pakwheels.com. The main part of this research work is “Introduction of suitable variable selection techniques, which helped to find that which variables are more

suitable and relevant for inclusion in model. This (His research) helps students and future researchers in many fields to understand the conditions under which studies should be conducted and gives them the knowledge to discern when appropriate techniques should be used [2].

Support Vector Machine (SVM) concept is used by one another researcher Mariana Listiani [3] for the same work. Listiani predicted prices of leased cars using above mentioned technique. It was found in this research that SVM technique is far better and more accurate for price prediction as compared to other like multiple linear regression when a very large data set is available. The researcher also showed that SVM also handles high dimensional data better and avoids both the under-fitting and over-fitting issues. To find important features for SVM Listiani used Genetic Algorithm. However, the technique failed to show in terms of variance and mean standard deviation because SVM is better than simple multiple regression [3].

Neural Networks (NN) are better in estimating price of house, this was concluded in the research of Limsombunchai[4]. By comparing with hedonic method his method was more accurate. Operation of both the methods are same, but in NN the model is trained first and then tested for prediction. Using both the methods NN produced higher R-sq and smaller root mean square error (RMSE), while hedonic produced lower values. This research was limited because the actual house price was missing, and only estimated prices were used for the research work [4].

K Noor and Saddaqat J [5] also worked to predict the price of Vehicles using different techniques. The researchers achieved highest accuracy using multiple linear regression. This paper proposes a system where price is dependent variable which is predicted, and this price is derived from factors like vehicle's model, make, city, version, color, mileage, alloy rims and power steering [5].

III. Methodology

The methodology used in this project involved several steps to build and evaluate the Mobile Phone Price Prediction model. The following is a detailed description of the methodology:

A. *Data Collection:* The data for our Mobile Phone Price Prediction project was collected through web scraping from the website "www.91mobiles.com". We targeted this specific website as it provides comprehensive information about various mobile phone models and their specifications. The data collection process involved accessing the website's pages, extracting the relevant information, and storing it in a structured format for further analysis. We sent HTTP requests to the website's URLs and retrieved the HTML content of the pages. Then, using BeautifulSoup, we parsed the HTML and extracted the required information such as the mobile phone's name, brand, RAM, internal memory, processor, camera specifications, battery capacity, Android version, UI, display type, resolution, thickness, weight, expandable memory, display size, refresh rate, and price. We iterated through multiple pages of the website to collect a substantial amount of data, ensuring diversity in terms of brands, models, and specifications. Each mobile phone record was stored as a row in our dataset, with each attribute mapped to its respective column. The collected data enabled us to analyze the relationships between various features and the corresponding prices of mobile phones, allowing us to build predictive models for price estimation.

RAM	RAM	373	object
Internal_Memory	Internal_Memory	373	object
Processor	Processor	373	object
back_camera	back_camera	373	object
front_camera	front_camera	372	object
battery	battery	372	object
android_version	android_version	373	object
UI	UI	373	object
Display_type	Display_type	373	object
Resolution	Resolution	373	object
Thickness	Thickness	373	object
weight	weight	373	float64

Fig 1: NON NULL Value Count of Features

B. *Data Preprocessing:* The dataset underwent several preprocessing steps, including handling missing values, converting data types, and cleaning the data. Categorical variables were encoded using one-hot encoding or label encoding, while numerical variables were scaled or normalized as needed.

Features	Minimum	Maximum	Mean	StdDiv
RAM	2	16	6.45699	2.23866
Internal_Memory	32	256	121.806	54.4755
back_camera	8	200	56.0484	30.411
front_camera	5	60	17.0968	10.5663
battery	2227	7000	4833.55	576.078
android_version	9	16	12.1287	1.09474
Thickness	6.3	10.08	8.29885	0.595208
weight	133	263	192.202	15.6598
display_size	5.4	7.6	6.56565	0.201166
Refresh_Rate	60	2000	506.321	394.892
price	5549	127999	28620.1	23390.3

Fig 2: Distinct Feature Information

C. *Feature Selection:* SelectKBest with chi-square test was used to select the top features that have the most significant impact on predicting the mobile phone price. The 10 best features were chosen based on their scores.

	Feature	Score
0	Brand	160.344285
1	RAM	139.901011
2	Internal_Memory	3829.335665
3	Processor	431.333286
4	back_camera	1084.925101
5	front_camera	754.206872
6	battery	8779.450999
7	android_version	12.370422
8	UI	68.494007
9	Display_type	35.827955
10	Resolution	1006.098978
11	Thickness	4.425810
12	weight	110.372229
13	display_size	0.130952
14	Refresh_Rate	24863.281030

Fig 3: Distinct Feature Information

D. *Model Training:* Two models were trained on the dataset: Random Forest Regression and Support Vector Regression. The Random Forest Regression model uses an ensemble of decision trees to make predictions, while the Support Vector Regression model utilizes support vector machines for regression tasks.

E. *Model Evaluation:* The trained models were evaluated using various evaluation metrics such as mean squared error (MSE). The MSE measures the average squared difference between the predicted and actual values, providing an indication of how well the models perform.

F. *Model Comparison:* The performance of the Random Forest Regression and Support Vector Regression models was compared based on their training and testing scores. The scores reflect the accuracy of the models in capturing the patterns and making predictions.

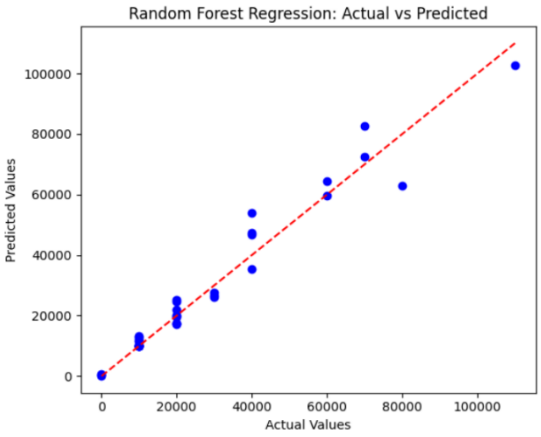

Fig 4: Actual vs Predicted Random Forest

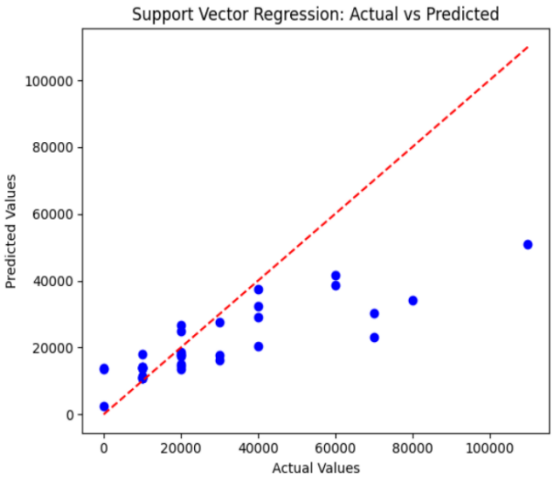

Fig 5: Actual vs Predicted SVM

G. *Data Visualization:* Data visualization techniques such as line plots and scatter plots were used to visually analyze the relationships between specific features (e.g., brand, RAM, internal memory, etc.) and the mobile phone prices.

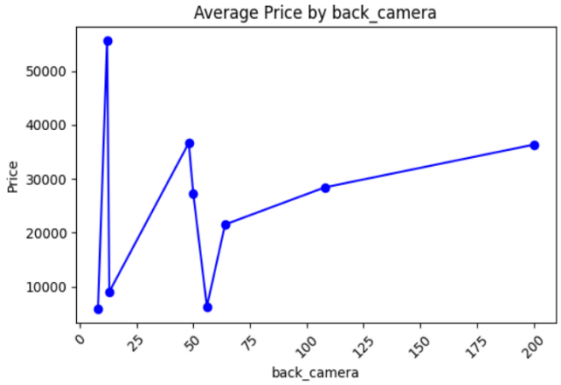

Fig 6: Average Price vs Back Camera

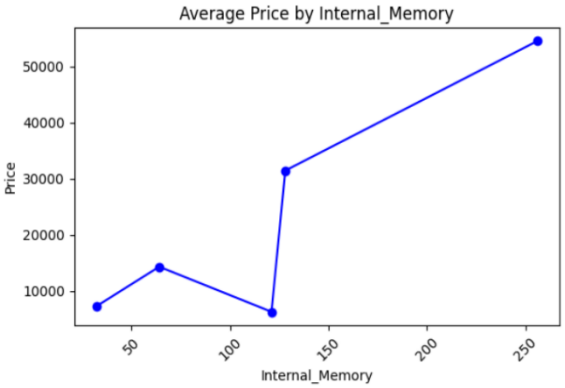

Fig 7: Average Price vs Back Camera

H. *Prediction:* The trained models were used to make predictions on new, unseen data, allowing for the estimation of mobile phone prices based on their features.

By following this methodology, the project aimed to develop an accurate and reliable Mobile Phone Price Prediction model using machine learning techniques. The next section will discuss the results and provide further insights into the performance and predictive capabilities of the models.

IV. Results

In this section, we present the results obtained from our Mobile Phone Price Prediction project using the Random Forest Regression and Support Vector Regression (SVR) models. We evaluated the performance of these models based on their mean squared error (MSE) values, as well as their training and testing scores.

Random Forest Regression Model
Mean Squared Error: 0.306
Training Score: 98.88%
Testing Score: 95.66%
The Random Forest Regression model demonstrated excellent performance, achieving a low mean squared error of 0.306. This indicates that the model's predictions were very close to the actual prices of mobile phones. The high training score of 98.88% suggests that the model learned the patterns and relationships in the training data effectively. Additionally, the testing score of 95.66% indicates that the model generalized well to unseen data, providing accurate price predictions for new mobile phone instances.

Support Vector Regression Model
Mean Squared Error: 3.108
Training Score: 24.25%
Testing Score: 47.87%
The Support Vector Regression model exhibited a relatively higher mean squared error of 3.108, suggesting that its predictions deviated further from the actual prices compared to the Random Forest model. The lower training score of 24.25% indicates that the model had difficulty capturing the underlying patterns in the training data. Similarly, the testing score of 47.87% indicates a moderate level of accuracy when predicting prices for new mobile phone instances.

Based on these results, we can conclude that the Random Forest Regression model outperformed the Support Vector Regression model in terms of both MSE values and training/testing scores. The Random Forest model demonstrated better accuracy and predictive performance for our Mobile Phone Price Prediction task.

V. Conclusion

In this project, we employed two machine learning algorithms, namely Random Forest Regression and Support Vector Regression, to train predictive models on the dataset. The Random Forest Regression model achieved a high training score of 98.88% and a testing score of 95.66%, indicating its effectiveness in predicting mobile phone prices. On the other hand, the Support Vector Regression model had lower training and testing scores of 24.25% and 47.87% respectively, suggesting that it may not be as suitable for this specific task.

The Random Forest Regression model achieved a lower MSE of 0.31, indicating its better performance in predicting prices compared to the Support Vector Regression model with an MSE of 3.11.

Overall, this project demonstrates the potential of machine learning techniques in predicting mobile phone prices. The findings can be valuable for consumers, retailers, and manufacturers in understanding the factors influencing mobile phone prices and making informed decisions in the mobile phone market.

VI. Future Work Extension

More sophisticated artificial intelligence techniques can be used to maximize the accuracy and predict the accurate price of the products. Software or Mobile app can be developed that will predict the market price of any newly launched product. To achieve maximum accuracy and predict more accurately, more and more instances should be added to the data set. And selecting more appropriate features can also increase the accuracy. So data set should be large and more appropriate features should be selected to achieve higher accuracy.

References

[1] Sameerchand Pudaruth. “Predicting the Price of Use Cars using Machine Learning Techniques”, International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 7 (2014), pp. 753 – 764

[2] Shonda Kuiper, “Introduction to Multiple Regression How Much Is Your Car Worth? ”, Journal of Statistics Education · November 2008

[3] Mariana Listiani , 2009. “Support Vector Regression Analysis for Price Prediction in a Car Leasing Application”. Master Thesis. Hamburg University of Technology.

[4] Limsombunchai, V. 2004. “House Price Prediction: Hedonic Price Model vs. Artificial Neural Network”, New Zealand Agricultural and Resource Economics Society Conference, New Zealand, pp. 25-26. 2004

[5] Kanwal Noor and Sadaqat Jan, “Vehicle Price Prediction System using Machine Learning Techniques”, International Journal of Computer Applications (0975 –8887) Volume 167 – No.9, June 2017.

[6] Mobile data and specifications online available from https://www.gsmarena.com/

[7] Introduction to dimensionality reduction, A computer science portal for Geeks.

[8] Ethem Alpaydın, 2004. Introduction to Machine Learning, Third Edition. The MIT Press Cambridge, Massachusetts London, England

[9] InfoGainAttributeEval-Weka Online available from election / InfoGainAttributeEval.html

[10] Thu Zar Phyu, Nyein Nyein Oo. Performance Comparison of Feature Selection Methods. MATEC Web of Conferences42 (2016).

Link to Solution

https://github.com/sahithreddy54321/Intel_Unnati_TheStinsons/tree/main