IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Phone Price Prediction Using ML Techniques

[1] G.Chamundeswari, [2] K. Srihari Teja, [3] Gorle Vasu, [4] K.Uma Gayatri, [5] K.Vikas
SIR C R Reddy College of Engineering, Eluru, Andhra Pradesh, India

Abstract—**The market for mobile phones in india is highly competitive, with new models being introduced frequently. Consumers are always looking for the latest technology and features, and they are willing to pay a premium price for it. In this paper, we propose a machine learning-based approach to predict the price of mobile phones based on various features such as brand, model, screen type, camera quality, and battery life. A dataset of mobile phones are collected with their corresponding features and prices from various online retailers. The data is processed and then different machine learning algorithms are applied such as linear regression, decision trees, and random forests to predict the price of mobile phones. The performance of the algorithms are evaluated using metrics such as root mean squared error, R-squared value and mean absolute error. Finally a model is selected based on the performance. The proposed system will help rural Indians in purchasing phones within their budget and with optimal specifications.**

Keywords— preprocessing, machine learning, metrics, regression, prediction

I.INTRODUCTION

Mobile phones have become an integral part of our daily lives, providing us with a wide range of functionalities such as communication, entertainment, and information access. With the rapid advancement of technology, mobile phones are now equipped with advanced features such as high-quality cameras, powerful processors, and long battery life. As a result, the market for mobile phones has become highly competitive, with numerous brands and models available in the market. Consumers are always looking for the latest technology and features, and they are willing to pay a premium price for it. Therefore, it is essential for manufacturers to price their products competitively to stay ahead of the competition. Accurately predicting the price of mobile phones can help manufacturers make informed decisions about pricing and marketing strategies. Similarly, consumers can benefit from accurate price predictions to make informed decisions when purchasing a mobile phone.

II.LITERATURE SURVEY

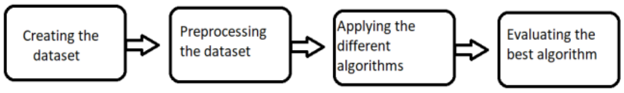
[6] stated that considering only strong attributes in the dataset gives more accurate results in random forest. [11] showed comparison between random forest and linear regression and concluded random forest as the best algorithm . [10] highlighted the significance of regression analysis in building ML models and its potential in practical applications. By providing a comprehensive overview of the main concepts and techniques used in regression analysis, this research has contributed to the development of the field and provided a foundation for further research in this area.

[7] contributed to the existing literature on modelling the non-linear complex behavior of suspended sediment responses to rainfall, water depth, and discharge in small catchment areas. The study highlights the importance of selecting the appropriate model and number of independent variables for suspended sediment discharge prediction. Overall, this paper demonstrates the potential of MLRg, MLP (LM, SCG, and BFGS), and RBF models in predicting suspended sediment discharge and emphasizes the importance of careful model selection and data preparation to achieve accurate results. The research contributes to the development of the field of environmental modelling and provides a foundation for further research in this area.[9] explores the use of correlation and linear regression techniques to test the relationship between two variables. Correlation measures the strength of the linear relationship between a pair of variables, while linear regression expresses the relationship in the form of an equation. Using simple examples and software tools such as SPSS and Excel, the article provides an overview of linear regression analysis and encourages readers to apply these techniques to their own data. The study highlights the importance of selecting the appropriate statistical technique based on the research question and type of data being analyzed. **[4**] emphasizes the importance of selecting the appropriate regression method based on the research question and data type. It highlights the advantages and limitations of each method and provides a valuable resource for researchers and data analysts.

[13] presents a new approach for solving the issue of high dependency among explanatory variables in regression analysis. The proposed approach is based on a ridge estimator, which is applied to study the relationship between macroeconomic variables and stock market movement. The results obtained from the proposed method are compared with those obtained from the ordinary least squares (OLS) method and it is observed that both methods provide similar results. The study concludes that the proposed method of estimation is capable of producing consistent results in the presence of multicollinearity in the data. [5] supports the effectiveness of random forests as a powerful and versatile tool for prediction in a variety of applications

III. PROPOSED SYSTEM

In the present market to buy a mobile we have to see either youtube reviews or reviews from different websites. There is no such system which predicts the price of phone based on their specifications. This model will predict the price of the phone based on its specifications. The proposed system will help rural Indians in purchasing phones within their budget and with optimal specifications.


Creating the dataset
Preprocessing the dataset
Applying the different algorithms
Evaluating the best algorithm

A. Creating the dataset

As the required dataset is not available anywhere, the dataset is manually populated by taking the price and different specifications of a particular mobile. This dataset contains different mobile specifications like brand, processor brand, processor model, screen type, battery, RAM, front camera. By considering all the specifications the price of the mobile is predicted.

B. Pre-processing the dataset

After taking all the specifications the dataset is pre-processed such that all the string and the large numerical values are changed to small numerical values by using the normalization and

C. Applying the different algorithms

After preprocessing the dataset different regression techniques are applied over the dataset for the prediction of the price. These different algorithms are trained over the dataset and the best among them is taken out.

D. Evaluating the best algorithm

After applying the different algorithms the best algorithm has to be taken out. For that process different regression evaluating methods like mse, mae, r square are used. By using these methods the best algorithm has been taken out.

Regression technique is used to find the price of the phone. Different regression techniques are used to find the price. All those techniques are evaluated by means of finding their metrics. Among all those techniques one technique is finalized based on its metrics. Different techniques used here are

- Linear Regression
- Multiple Linear Regression
- Polynomial Regression
- Support Vector Regressor
- Decision Tree
- Random Forest
- Lasso Regression
- Ridge Regression1
- Bayesian Ridge
- PLS Regression
- Elastic Net Regression

Among all these techniques one technique selected based on their performances.

1. Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. In simple linear regression, there is only one independent variable, and the goal is to fit a straight line (referred to as the "regression line") that best predicts the dependent variable based on this independent variable. The regression line is represented mathematically as an equation, and the coefficients in this equation are estimated from the data using a technique such as least squares. Simple linear regression can be used for both continuous and binary dependent variables.

The simple linear regression formula is given by

y = b0 + b1 * x

y is the dependent variable (the one being predicted)

x is the independent variable (the one used to make predictions)

b0 is the y-intercept (the value of y when x = 0)

b1 is the slope (the change in y corresponding to a change of 1 unit in x)Where T_{Wg} and K_{Wg} are the time constant and gain constants of wind turbine generator respectively.

2. Multiple Linear Regression

Multiple linear regression is a machine learning method used to model the relationship between a dependent variable (y) and multiple independent variables (x1, x2, ..., xn). The formula for multiple linear regression is:

y = b0 + b1 * x1 + b2 * x2 + ... + bn * xn

where y is the dependent variable (the one being predicted). x1, x2, ..., xn are the independent variables (the ones used to make predictions)

b0 is the y-intercept (the value of y when all independent variables are equal to 0). b1, b2, ..., bn are the coefficients representing the effect of each independent variable on the dependent variable. These coefficients must be estimated from the data.

3. Polynomial Regression

Polynomial Regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modeled as an nth degree polynomial. It is a form of linear regression, but instead of fitting a straight line to the data, we fit a polynomial curve of degree n. The equation of the polynomial curve can be represented as:

y = β0 + β1x + β2x^2 + ... + βnx^n

where β0, β1, β2, ..., βn are the coefficients of the polynomial, and n is the degree of the polynomial. The coefficients can be estimated using optimization algorithms such as gradient descent or least squares.

4. Support Vector Regressor

Support Vector Regression (SVR) is a type of supervised machine learning algorithm that is used for regression tasks (predicting a continuous outcome variable). It is a modification of the Support Vector Machine (SVM) algorithm, which is primarily used for classification tasks. SVR aims to find the optimal line or hyperplane that best separates the data points in the feature space, so as to minimize the error in prediction. The algorithm uses a kernel function to map the input data into a higher-dimensional space, making it possible to handle non-linear relationships between the independent and dependent variables. SVR has been used in a wide range of applications, including financial forecasting, sales forecasting, and engineering design.

$$f(x)=\beta_0+ \Sigma_{i=1}^n \alpha_i y_i K(x_i, x)$$

$f(x)$ is the predicted output for a given input x

β_0 is the bias term

α_i are the weights or coefficients that determine the influence of each training example on the prediction. y_i are the target values for each training example. x_i are the input feature values for each training example

$K(x_i, x)$ is a kernel function that maps the input data into a higher-dimensional space

5. Decision Tree

A Decision Tree is a tree-based model used in decision analysis, machine learning and statistics to predict a target variable by learning simple decision rules inferred from the data features. At each node of the tree, a decision rule is formed to split the data based on the value of a feature that results in the largest reduction in impurity (e.g. Gini impurity, entropy) of the target variable. This process continues recursively for each resulting subgroup of the data until a stopping criteria is met (e.g. minimum sample size, maximum tree depth).

6. Random Forest

Random Forest is a machine learning algorithm that is used for classification and regression. It is a collection of multiple decision trees, where each tree is trained on a random subset of the data and outputs a prediction. The final prediction of a Random Forest is determined by taking the average (for regression) or majority vote (for classification) of the predictions from individual trees. This combination of multiple trees makes the algorithm more robust to overfitting compared to a single decision tree.

7. Lasso Regression

Lasso Regression is a type of regularized linear regression algorithm that uses L1 regularization. Lasso stands for Least Absolute Shrinkage and Selection Operator.

The formula for Lasso Regression is as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

where:

y is the target/dependent variable

$x_1, x_2, \ldots, x_n$ are the independent/predictor variables

$\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are the regression coefficients

ε is the error term (representing the difference between the actual and predicted values)

8. Ridge Regression

Ridge Regression is a type of linear regression that is regularized using the L2 norm of the coefficients. It is used to prevent overfitting in linear regression models by adding a penalty term to the loss function that discourages large coefficients.The regularization term in Ridge Regression is the sum of the squared coefficients, multiplied by a regularization strength or hyperparameter, alpha. The larger the value of alpha, the stronger the regularization and the smaller the magnitude of the coefficients.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

where:

y is the target/dependent variable

$x_1, x_2, \ldots, x_n$ are the independent/predictor variables

$\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are the regression coefficients

ε is the error term (representing the difference between the actual and predicted values)

9.Bayesian Ridge

In Bayesian Ridge Regression, the coefficients are estimated as the maximum a posteriori (MAP) estimates, which are the values of the coefficients that maximize the posterior distribution given the data and prior knowledge. The MAP estimates are found by using an optimization algorithm, such as gradient descent, to minimize the negative log-posterior probability of the coefficients.

$$p(\beta \mid X, y) = p(y \mid X, \beta) * p(\beta) / p(y \mid X)$$

where:

β is the vector of coefficients

X is the matrix of predictors

y is the target variable

$p(y \mid X, \beta)$ is the likelihood function that models the relationship between the predictors and the target variable

$p(\beta)$ is the prior distribution on the coefficients, which represents prior knowledge about the relationships between the predictors and the target variable

$p(y \mid X)$ is the marginal likelihood or evidence, which is the normalizing constant for the posterior distribution

Once the MAP estimates have been found, the prediction for a new observation can be expressed as:

$$y\_pred = X\_new * \beta\_map$$

10. Elastic Net Regression

ElasticNet is a regularization technique in machine learning, which combines the L1 (Lasso) and L2 (Ridge) regularization methods to balance the strength of the penalties imposed on the coefficients of the model.In Lasso regularization, the absolute values of the coefficients are penalized, leading to sparse solutions with many coefficients set to zero. In Ridge regularization, the squared values of the coefficients are penalized, leading to solutions with smaller, non-zero coefficients. ElasticNet combines these two regularization methods by adding both L1 and L2 penalties to the loss function. The combination of L1 and L2 regularization in ElasticNet allows for a trade-off between sparsity and shrinkage, providing a solution that is more flexible and less likely to overfit the data compared to either Lasso or Ridge regularization alone. The amount of L1 and L2 regularization can be controlled through a mixing parameter, which determines the relative strength of each penalty.

Loss = (1/n) * SUM(y - y_pred)^2 + α * (ρ * L1_penalty + (1 - ρ) * L2_penalty)

where:

n is the number of observations

y is the true target value

y_pred is the predicted target value

α is the regularization strength or hyperparameter that controls the overall magnitude of the penalties

ρ is the mixing parameter that determines the relative strength of the L1 and L2 penalties

L1_penalty = SUM(|β|) is the absolute sum of the coefficients

L2_penalty = SUM(β^2) is the squared sum of the coefficients

y_pred is the predicted target value

α is the regularization strength or hyperparameter that controls the overall magnitude of the penalties

ρ is the mixing parameter that determines the relative strength of the L1 and L2 penalties

L1_penalty = SUM(|β|) is the absolute sum of the coefficients

L2_penalty = SUM(β^2) is the squared sum of the coefficients

11.PLS Regression

PLSR stands for Partial Least Squares Regression, which is a statistical technique used for modelling the relationship between a set of independent variables and a set of dependent variables. It is a type of regression analysis that helps to reduce the dimensionality of the data while maintaining the maximum amount of information PLSR is commonly used in fields such as chemometrics and genomics, where the number of predictor variables is large compared to the number of observations. The formula for PLSR involves a series of regression equations, which are used to estimate the weights or loadings for each independent variable, as well as the regression coefficients for the dependent variables.

Mathematically, PLSR can be expressed as follows:

X = TP' + E, where X is the matrix of independent variables, T is the matrix of scores (weights), P is the matrix of loadings (regression coefficients), and E is the residual matrix.

Y = UQ' + F, where Y is the matrix of dependent variables, U is the matrix of scores (weights), Q is the matrix of loadings (regression coefficients), and F is the residual matrix.

The scores and loadings are estimated by iteratively minimizing the residuals in both X and Y. The final regression equation can then be expressed as Y = XB + E, where B is the matrix of regression coefficients.

IV. EXPERIMENTAL RESULTS

At this stage the metric for the different algorithms are found. Based on those metrics the best algorithm is evaluated for the prediction of the price. For the calculation of the metric the used methods are

- R Square
- Mean Absolute Error
- Mean Square Error

a. R Square Method

R-squared (R^2) is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variable(s) in a regression model.

Formula: R^2 = 1 - (sum of squares of residuals) / (sum of squares of total variance in the dependent variable)

Where the residuals are the differences between the actual values of the dependent variable and the predicted values from the regression model. The total variance in the dependent variable is calculated as the sum of squares of the differences between the actual values and the mean of the dependent variable.

R^2 is a value between 0 and 1, where 0 indicates that the model does not explain any variance in the dependent variable, and 1 indicates that the model explains all the variance in the dependent variable.

TABLE.I.PERFORMANCE OF DIFFERENT TECHNIQUES BY R2 SCORE

Regression Technique	R2 Score
Simple Linear Regression	0.358
Multiple Linear Regression	0.358
Support Vector Regressor	0.254
Polynomial	0.0712
Decision tree	0.369
Random forest	0.6157
Lasso Regression	0.5786
Ridge Regression	0.5786
Bayesian ridge	0.5147
Elastic Net Regression	0.5783
PLS Regression	0.5789

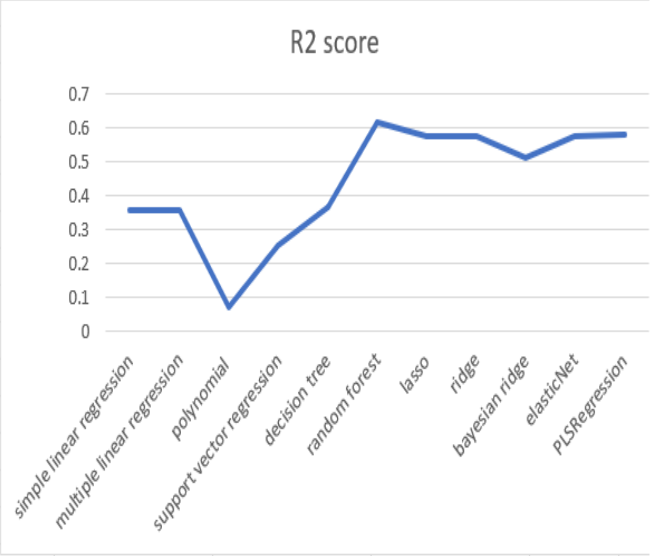


Fig. 1.Graph of R2 Score

b. Mean Absolute Error

Mean Absolute Error (MAE) is a measure of the difference between the actual and predicted values in regression problems. It is the average absolute difference between the predictions and actual values.

Formula: MAE = (1/n) * Σ|actual_i - predicted_i|

Where n is the number of samples, actual_i is the actual value of the i-th sample, and predicted_i is the predicted value of the i-th sample.

MAE is a robust measure of error that is not sensitive to extreme values, unlike Mean Squared Error (MSE). It provides a more interpretable output as the error values are expressed in the same unit as the target variable, making it easier to understand the magnitude of the error.

TABLE. II.PERFORMANCE OF DIFFERENT TECHNIQUES BY MEAN ABSOLUTE ERROR

Regression Technique	Mean Absolute Error
Simple Linear Regression	6794.223789
Multiple Linear Regression	6146.553891
Support Vector Regressor	7702.347107
Polynomial	5780.311673
Decision tree	6602.753981
Random forest	4189.89461
Lasso Regression	5704.342123
Ridge Regression	5114.130304
Bayesian ridge	6935.785901
Elastic Net Regression	4788.073986
PLS Regression	5780.542138

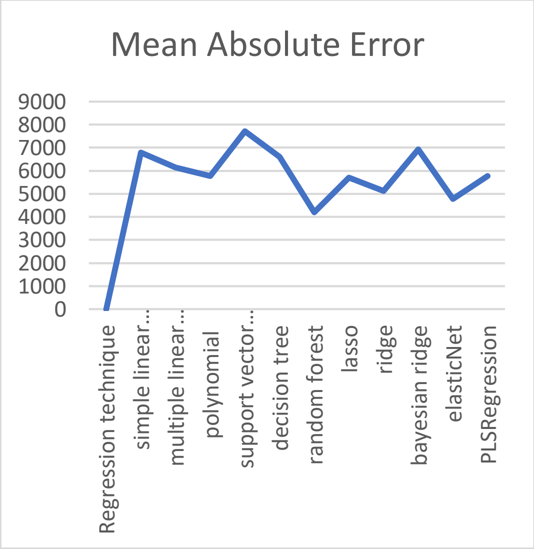


Fig. 2 .Graph of Mean Absolute Error

c. Mean Square Error

Mean Squared Error (MSE) is a commonly used measure of the difference between the predicted values and the true values. It represents the average of the squared differences between the predictions and the actual values.

The formula for MSE is:

MSE = (1/n) * Σ(predicted value - actual value)^2

Where n is the number of observations and Σ is the sum across all observations.

TABLE. III.PERFORMANCE OF DIFFERENT TECHNIQUES BY MEAN SQUARE ERROR

Regression Technique	Mean Absolute Error
Simple Linear Regression	88822517.85
Multiple Linear Regression	54940977.41
Support Vector Regressor	106822909.53
Polynomial	54675831.03
Decision tree	128835471.94
Random forest	37772002.96
Lasso Regression	54364202.25
Ridge Regression	52890319.96
Bayesian ridge	76832984.03
Elastic Net Regression	40451989.88
PLS Regression	54680414.67

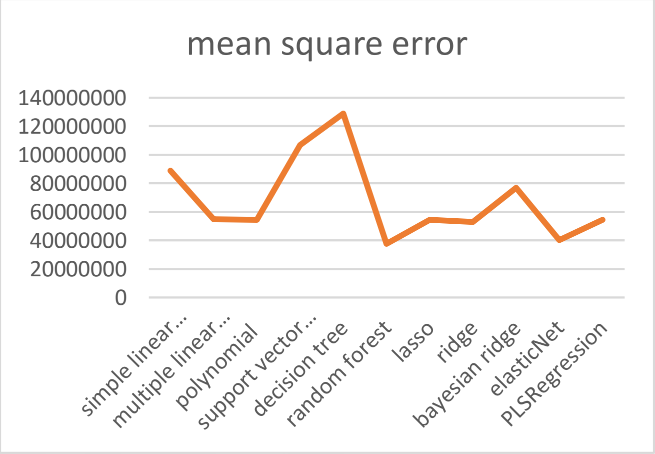


Fig. 3.Graph of Mean Square Error

VI. CONCLUSION

This research paper reflects the metric evaluation of different regression techniques on the same dataset. Among all those different regression techniques Random Forest came as the best technique. By this metric evaluation the conclusion is that the Random Forest is best suitable for the prediction of the price. After taking the Random Forest as the best, based on the metrics. Then using the streamlit, interface is created which is based on the random forest. The price is predicted based on the specification.

REFERENCES

[1] Basak, Debasish & Pal, Srimanta&Patranabis, Dipak. (2007). Support Vector Regression. Neural Information Processing – Letters and Reviews. 11.

[2] Ali, Jehad & Khan, Rehanullah& Ahmad, Nasir & Maqsood, Imran. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues(IJCSI). 9.

[3] Kumari, Khushbu & Yadav, Suniti. (2018). Linear regression analysis study. Journal of the Practice of Cardiovascular Sciences. 4. 33. 10.4103/jpcs.jpcs_8_18.

[4] Sarstedt, Marko & Mooi, Erik. (2014). Regression Analysis. 10.1007/978-3-642-53965-7_7.

[5] Ali, Jehad & Khan, Rehanullah& Ahmad, Nasir & Maqsood, Imran. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues(IJCSI). 9. Analysis of a Random Forests Model

[6] Uca, &Toriman, Mohd & Jaafar, Othman & Maru, Rosmini&Arfan, Amal & Ahmar, Ansari. (2018). Daily Suspended Sediment Discharge

Prediction Using Multiple Linear Regression and Artificial Neural Network. Journal of Physics: Conference Series. 954. 012030. 10.1088/1742-6596/954/1/012030.

[7] Maurice, Wanyonyi. (2020). Modelling Factors Affecting Lung Capacity. Journal of Advances in Mathematics and Computer Science. 1-18. 10.9734/jamcs/2019/v34i/630229.

[8] Kumari, Khushbu & Yadav, Suniti. (2018). Linear regression analysis study. Journal of the Practice of Cardiovascular Sciences. 4. 33. 10.4103/jpcs.jpcs_8_18.

[9] Langley, Pat &Kibler, Dennis. (1997). The Experimental Study of Machine Learning.

[10] Jui, Julakha& Molla, M. M. Imran & Bari, Bifta& Rashid, Mamunur& Hasan, Md Jahid. (2020). Flat Price Prediction Using Linear and Random Forest Regression Based on Machine Learning Techniques. 10.1007/978-981-15-6025-5_19.

[11] Kwon, Sunghoon& Han, Sangmi& Lee, Sangin. (2013). A small review and further studies on the LASSO. Journal of the Korean Data and Information Science Society. 24. 10.7465/jkdi.2013.24.5.1077.

[12] Mohamad Shariff, Nurul Sima&Duzan, H. (2018). An Application of Proposed Ridge Regression Methods to Real Data Problem. International Journal of Engineering & Technology. 7. 106. 10.14419/ijet.v7i4.30.22061.