Mobile Price Prediction

Introduction -

Mobile price prediction is an important task in the field of mobile commerce, helping both consumers and businesses make informed decisions. Researchers have explored various approaches and techniques to predict mobile prices accurately. This report presents a comprehensive study on mobile price prediction using machine learning techniques which can have practical implications for consumers, manufacturers, and market analysts in making informed decisions and enhancing pricing strategies. We explore various features, such as hardware specifications, brand, camera quality, and battery capacity, to develop accurate models for predicting mobile prices.

Literature Survey -

Machine Learning-Based Approaches:
Many researchers have employed machine learning techniques to predict mobile prices based on various features and attributes. They have utilized algorithms such as linear regression, decision trees, random forests, support vector machines (SVM), and neural networks as used in [1] and [2]. These approaches aim to capture the relationships between the mobile phone features (e.g., brand, specifications, performance, camera quality, etc.) and their corresponding prices. By training models on historical mobile price data, these approaches can predict prices for new mobile phones accurately[3]. As for now there are many resources for mobile price range prediction and very less resources for actual mobile price prediction.

Feature Selection and Importance:
Feature selection plays a crucial role in mobile price prediction. Researchers have conducted studies to identify the most influential features that contribute significantly to the price determination. They have employed feature selection algorithms like correlation analysis, information gain, and recursive feature elimination to identify the relevant features. By considering only the most informative features, the prediction models can achieve better accuracy and efficiency.[1][2][3]

Data Preprocessing Techniques:
Preprocessing the mobile price dataset is crucial for achieving accurate predictions. Researchers have employed various techniques such as data cleaning, normalization, and feature scaling. Data cleaning involves handling missing values, outliers, and inconsistencies in the dataset. Normalization techniques and Scaling techniques are used example Robust Scaler or Standard Scaler. These preprocessing techniques ensure that the input data is in a suitable format for training the prediction models.[3][4]

Comparative Studies and Evaluation Metrics:
Researchers have compared the accuracy, precision, recall, F1-score, mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R-squared (R2) as evaluation metrics for price prediction as in [4]. These studies help in identifying the most effective approaches and models for mobile price prediction tasks. For my project main evaluation metric will be R-squared score and MSE.

Methodology -

Data Collection:

I have taken the dataset from Kaggle for this project. The mobile price prediction dataset contains information about different mobile phones and their corresponding prices. The dataset includes features such as brand, RAM (Random Access Memory), internal memory, battery capacity, camera features, screen size, resolution, and other specifications. Additionally, the dataset provides the price of each mobile phone, which serves as the target variable for prediction.

EDA:

EDA is a crucial step in understanding the dataset and gaining insights into the relationships between variables. It helps in identifying patterns, trends, and potential outliers in the data. In the context of mobile price prediction, EDA involves analyzing both numerical and categorical variables, as well as exploring their relationships with the target variable (price). Univariate analysis involved exploring individual variables to understand their distributions, while bivariate analysis focused on examining relationships between variables and the target variable (price) in the context of mobile price prediction. Descriptive statistics, density plots, scatter plots, and bar plots were used to analyze the data and identify patterns and correlations. These analyses provided insights into variable distributions and their impact on price prediction. (Explained in detail in notebook)

Feature Engineering and Selection :

Feature engineering is performed to extract additional meaningful information from the existing variables. In this case, a new feature called "PPI" (Pixels Per Inch) can be created by combining the resolution (x and y) and screen size variables. This feature represents the pixel density of the mobile screen and can potentially have an impact on the mobile price. [6][7] The PPI can be calculated by dividing the square root of product of resolution (x^2 and y^2) by the screen size. This new feature can then be included in the analysis and explored for its relationship with the price.

Feature selection technique are performed like chi-square test, SeleckKBest and correlation coefficient[1][8]. Chi-Square test measures the independence between a categorical feature and the target variable. By applying the chi-square test to each categorical feature, we can determine which features have a significant association with the target variable (price). For the mobile price prediction task, SelectKBest can be applied with an appropriate scoring function, such as chi-square for categorical variables or mutual information for numerical variables. The algorithm ranks the features based on their scores and selects the top K features with the highest scores. By calculating the correlation coefficients (e.g., Pearson correlation) between each numerical feature and the target variable (price), we can assess the strength and direction of the relationship. Features with higher absolute correlation coefficients (positive or negative) are more likely to be important predictors of mobile prices.

After these feature selection steps all the unnecessary features like 'Touchscreen', 'Wi-Fi', 'Bluetooth', 'Model', ‘Screen Size’,’Resolution (x and y)’ are dropped from the dataset. By combining feature engineering techniques such as creating new features (e.g., PPI) and utilizing feature selection methods like chi-square test, SelectKBest, and correlation analysis, we can identify the most relevant features that contribute significantly to mobile price prediction. These selected features can improve the model's

performance and lead to more accurate price predictions.

Data Preprocessing: The entire dataset is split into train and test set with a generic 80-20 split. Our main focus will be to evaluate different model and see the metric on the test set basically how well our model is predicting price based on the R2 score and MSE. All the models are implemented in a pipeline method [9]. First all the numerical features are scaled using Robust Scaler and the Categorical Features are one - hot encoded. In the next step of pipeline model is applied and then we fit the model to our dataset and the predict price values.

Metric :

In the context of mobile price prediction, the selection of appropriate evaluation metrics and models is crucial for assessing the performance and accuracy of the predictions. In this case, the chosen evaluation metrics are the coefficient of determination (R2 score) and mean squared error (MSE). [3][4]

R2 score - The R2 score measures the proportion of the variance in the target variable that is explained by the model. It is calculated as 1 - (sum of squared residuals / total sum of squares).

MSE (Mean Squared Error): It is calculated as the average of the squared differences between predicted and actual values.

The decision to use the R2 score and MSE as evaluation metrics stems from their interpretability and ability to capture different aspects of model performance. The R2 score provides an overall assessment of how well the model explains the variance in mobile prices, while the MSE focuses on the magnitude of prediction errors. By selecting these metrics, we can gain insights into the model's explanatory power and the accuracy of price predictions.

Model Selection :

Baseline Model: Basic Linear Regression without log transformation of prices

The equation of Linear Regression is:

$y = \beta0 + \beta1x1 + \beta2x2 + ... + \beta n*xn$

where:

y is the predicted value of the target variable,

β0 is the y-intercept,

β1, β2, ..., βn are the coefficients of the independent variables (x1, x2, ..., xn),

x1, x2, ..., xn are the values of the independent variables.

(In all the models after this target price column is log transformed.)

First Model- Linear Regression with log transformation of prices because of the results of the baseline model.

Second Model - SVR

Equation: $y = b0 + b1x1 + b2x2 + ... + bn*xn$.

SVR is a regression model that uses support vector machines to find a hyperplane that best fits the data.

Main reason for using SVR after LR is to capture non-linear relationships and improve robustness to outliers, leveraging its kernel functions and loss function.

It is a non-linear type of regression model.

Third Model- Random Forest[10]

Random Forest Regressor is an ensemble learning method that uses a collection of decision trees to make predictions. It combines the predictions from multiple trees to produce a more accurate and robust result. Reason for using random forest is that they are less prone to overfitting compared to linear regression and can capture complex patterns in the data.

Fourth Model - Gradient Boosting

Gradient Boosting is an ensemble learning technique that combines multiple weak learners to create a strong predictive model. It sequentially trains models on the residuals of the previous models, improving the overall prediction by minimizing a loss function.Gradient

boosting was chosen after RF because it combines multiple weak learners to create a strong model by iteratively minimizing the residuals. This iterative process helps in reducing bias and improving overall accuracy, making it effective for predicting prices accurately.

Experimental Results (Evaluation) -

For visualization of result the histogram plot (histplot) is chosen to compare the actual and predicted values because it provides an intuitive way to visualize the distribution and overlap between the two sets of values. It allows for a quick assessment of how well the predicted values align with the actual values and provides insights into any potential discrepancies or patterns. The visualization can be seen in the notebook.

Baseline Model - (LR without log transformation)

Output -

```
MSE Train Loss 66870017.30473149
RMSE Test Loss  6881.610475078928
R2 score 0.48709550197095564
MAE 4514.521018430173
```

We can see that the R2 score is very less and RMSE loss is also very high the reason can be because of price skewed distribution. So we can log transform our price column as done in [4] which shows better results after log transformation of price column. By taking the log of prices, we can transform the data to achieve a more symmetric distribution and potentially capture non-linear relationships more effectively and also reduce outliers and heteroscedasticity.

First Model- Linear Regression

After log transformation of target price column and applying linear regression again we see improvement in the results.

```
MSE Train Loss: 0.143101748852161 62
MSE Test Loss: 0.14105584672191113
R2 Score: 0.700880657207523
MAE: 0.28682587478920585
```

The R2 score increased drastically and the MSE loss also very less for the basic model as explained before.

Second Model - SVR

Output -

```
MSE Train Loss: 0.11872114364721421
MSE Test Loss: 0.12352364177311297
R2 Score: 0.738058992908321
MAE: 0.2695520570466469
```

The R2 score increased and MSE Loss decreased showing the model is performing better than linear regression. SVR uses a kernel function to transform the data into a higher-dimensional space, allowing it to model complex patterns that cannot be effectively captured by linear regression.

Third Model - Random Forest

Output -

```
MSE Train Loss: 0.0733582270331224
MSE Test Loss: 0.15182585795858788
R2 Score: 0.6780420527976434
MAE: 0.2945780497917041
```

The R2 score decreased and MSE test loss also increased possible reason being Random Forest Regressor is a powerful model that can handle complex relationships and capture non-linearities in the data. However, in certain cases, such as when the data has a linear relationship or when there are only a few important features like in our dataset simpler models like SVR and Linear Regression can outperform Random Forest. Also one more possible reason can be the hyperparameters used in model training. We can try to tune the hyperparameters using GridSearchCV which can be done in future work.

Fourth Model - Gradient Boosting

Output -

```
MSE Train Loss: 0.049484123417731535
MSE Test Loss: 0.12210401361085538
R2 Score: 0.7410694192945548
MAE: 0.26216233939341566
```

The R2 score increased and the MSE loss decreased and is the best among all the 4 models. So the model which performed the best was this model as expected.

Future Work -

In future work, it would be beneficial to investigate the impact of including additional relevant features such as customer reviews, and market trends to further improve the accuracy of mobile price prediction models. Furthermore, incorporating time series analysis techniques and considering the dynamic nature of the mobile market could lead to more precise and up-to-date price predictions.

Refrence-

[1] Mobile Price Prediction Using Machine Learning Article Link

[2] A. Kalmaz and O. Akin, "Estimation of Mobile Phone Prices with Machine Learning," 2022 International Conference on Engineering and Emerging Technologies (ICEET), Kuala Lumpur, Malaysia, 2022, pp. 1-7, doi: 10.1109/ICEET56468.2022.10007128. Paper Link

[3] Chandrashekhara, K.T., Thungamani, M., Gireesh Babu, C.N., Manjunath, T.N. (2019). Smartphone Price Prediction in Retail Industry Using Machine Learning Techniques. In: Sridhar, V., Padma, M., Rao, K. (eds) Emerging Research in Electronics, Computer Science and Technology. Lecture Notes in Electrical Engineering, vol 545. Springer, Singapore. Paper Link

[4] Housing Price prediction Using Support Vector Regression Jiao Yang Wu San Jose State University PDF Link

[5] Book - Machine Learning Yearning BY AndrewNg

[6] Our Guide to PPI (Pixels per Inch) and PixelDensity. Link

[7] How Mobile Screen Size, Resolution, and PPI Screen Affect Test Coverage Link

[8] Feature Selection Techniques in Machine Learning (Updated 2023) Blog Link

[9] Pipelines – Python and scikit-learn Link

[10] Random Forest Regression Link

[11] Scikit Learn Documentation

[12] Seaborn Documentation