# Report on Smart Mobile Phone Price Prediction Using Machine Learning

By Team Alpha Debuggers

## 1. Literature Survey and Related Work

### 1.1 Importance of Mobile Phone Price Prediction

Mobile phone price prediction is important for several reasons:
-> **Customer Decision Making:** Price is one of the primary factors influencing customers' purchasing decisions. Predicting mobile phone prices accurately helps customers make informed decisions based on their budget and value for money.
-> **Market Competition:** Price also plays a crucial role in the competitive landscape of the mobile phone market. By accurately predicting prices, manufacturers and retailers can strategically position their products to gain a competitive edge, attract customers, and maximize profitability.
-> **Marketing and Sales Strategy:** Accurate price prediction enables companies to develop effective marketing and sales strategies. They can determine the optimal pricing strategy based on market demand, competition, and target customer segments. This helps in maximizing revenue and profitability while maintaining market share.

In summary, mobile phone price prediction is crucial for customer decision making, market competition. It is also important for marketing and sales strategy, inventory management, product development, financial planning, and market analysis. It enables companies to make informed decisions, optimize pricing strategies, and enhance overall business performance in the dynamic mobile phone market.

### 1.2. Past Works

Several studies have explored the prediction of mobile phone prices using machine learning techniques. Muhammad Asim and Zafar Khan from UET Lahore focused on converting the regression problem into a classification problem and employed decision trees and Naive Bayes algorithms with the WEKA software. Their dataset, collected from GSM Arena, included ten features to predict price classes. Another project by Nivitus as a blog on Medium involved web scraping from an e-commerce website, utilizing features such as brand, ratings, RAM, ROM, and others. The author achieved high accuracy (95-96%) using Random Forest Regressor and Support Vector Regression algorithms.

A study conducted by Ahsanul Hoque Sakib and supervised by Asif Khan Shakir at Daffodil University in Bangladesh used decision trees, Naive Bayes, and random forest algorithms to predict mobile price ranges. Their dataset from Kaggle consisted of 10 features, with RAM being identified as the most important feature in the decision tree algorithm. Prateek Majumdar discussed mobile price prediction in a blog on Analytics Vidyalaya using four classification algorithms: K-nearest neighbors (KNN), random forest, support vector machine (SVM), and Naive Bayes. The author implemented these algorithms in Python and reported successful prediction of mobile prices.

Overall, these studies showcased the application of various machine learning techniques in predicting mobile prices, with reported accuracies ranging from 83% to 96% for the prediction models. For our project, we plan to implement similar models in Python, although we will be using a different dataset.

**Flowchart of our project**:
https://drive.google.com/file/d/1Y19HZeE1eajMU5C9wZqdkM3PpAPbBHD6/view?usp=sharing

This is a basic workflow of the steps in our project from collecting the data from a source, to processing it into a table. Then we choose some good ML models & train it with the data. Finally we predict prices, evaluate the models & show our results.

## 2. Dataset Selection and Exploratory Data Analysis

We used a recent Kaggle dataset with 408 phones & 10 features. This dataset contains information on the prices of several mobile phones from different brands. It includes details such as the storage capacity, RAM, screen size, camera specifications, battery capacity, and price of each device. It can be considered a good dataset for analyzing and comparing different smartphone models based on these specifications.

- **Brand:** the manufacturer of the phone

- **Model:** the name of the phone model

- **Storage (GB):** the amount of storage space (in gigabytes) available on the phone

- **RAM (GB):** the amount of RAM (in gigabytes) available on the phone

- **Screen Size (inches):** the size of the phone's display screen (diagonal length) in inches

- **Camera (MP):** the megapixel count of all the phone's camera(s)

- **Battery Capacity (mAh):** the capacity of the phone's battery in milliampere hours

- **Price ($):** the retail price of the phone in US dollars

The provided dataset offers valuable insights into pricing trends and allows for a comprehensive analysis of the features and prices of various mobile phone models. Each row represents a different phone model, enabling comparisons and identification of patterns in pricing and specifications. Features such as brand, model, storage, RAM, screen size, camera, battery capacity, and price play significant roles in determining the overall user experience and suitability of a phone. Users can utilize this dataset to make informed decisions based on their preferences, needs, and budget, considering factors that matter most to them, such as brand reputation, storage capacity, camera quality, and price affordability. However, it is essential to note that the importance of these features may vary among individual consumers, as personal preferences and priorities differ.

For retailers, the dataset provides a valuable resource for pricing decisions and market analysis. Along with brand, model, and price, retailers may consider factors like market demand, consumer preferences, competition, and additional features when determining phone prices. Understanding the market dynamics and consumer behavior can help retailers identify pricing strategies that align with their target market's preferences and expectations. However, it's important to consider other factors like currency, data completeness, and relevance to ensure the accuracy and applicability of the dataset for specific retail scenarios.

In conclusion, the provided dataset serves as a solid foundation for analyzing pricing trends, comparing features, and understanding the relationship between specifications and prices of different mobile phone models. It offers insights for both individual consumers and retailers, allowing for informed decision-making and market analysis. However, users should consider the limitations of the dataset and account for additional factors specific to their needs and market conditions.

We have done Exploratory Data Analysis(EDA) in the code.The output of the code will show the first few rows of the dataset using (*pd.head()*), summary statistics (mean, standard deviation, minimum, maximum, quartiles) for numerical features(*pd.describe()*), and the count of missing values for each feature(*pd.isnull().sum()*). We graph line plots, scatter plots & bar graphs in Python to measure the importance of the matrics.

Based on this, we select Brand, Storage, RAM, Screen Size, Battery Capacity & Camera. These are some of the most important factors of selecting a phone. Other datasets have used extra features like Model, Weight, NFC (Near Field Communication), Clock Speed, Bluetooth and more but they are not as important to the price.

## 3) Metric & Model Selection

Some commonly used metrics for regression tasks include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (coefficient of determination).

MAE, RMSE, MSE, and R-squared are commonly used metrics in the field of statistics and machine learning to evaluate the performance of regression models.

In summary, MAE, RMSE, and MSE focus on the magnitude of the prediction errors, while R-squared measures the proportion of variance explained by the model. The choice of which metric to use depends on the specific requirements and characteristics of the problem at hand.

For mobile price prediction, metrics like MSE and RMSE are commonly used as they provide an indication of the average squared or square-rooted difference between the predicted and actual prices. These metrics give more weight to larger errors, which can be useful in assessing the overall accuracy of the model.

However, it's important to consider the specific context and requirements of the project. For example, if the goal is to identify outliers or focus on minimizing errors on high-priced phones, other metrics like Mean Absolute Error (MAE) or Median Absolute Error (MedAE) or percentage errors may be more suitable.

Based on previous work done in the field, we plan to select Linear Regression, Logistic Regression, KNN, SVR and Random Forest & test them all out & compare them.

Choosing multiple models to test and compare is a good approach to understand their performance and determine the most suitable one for our mobile price prediction task. We will briefly discuss each of the models we mentioned:

1) **Linear Regression:** It is a simple and interpretable model that assumes a linear relationship between the input features and the target variable. It can provide insights into the importance and direction of each feature's contribution to the price prediction. The model estimates the coefficients for each feature and uses them to make predictions. We can evaluate the model's performance using metrics such as MSE, RMSE, MAE, or R-squared.
2) **Logistic Regression:** It seems like we intended to include logistic regression, which is typically used for binary classification tasks. Since mobile price prediction is a regression task, logistic regression may not be the most appropriate choice. However, if we have a classification component (e.g., predicting whether a phone is low, medium, or high priced), logistic regression can be used for that aspect.
3) **K-Nearest Neighbors (KNN):** KNN is a non-parametric model that predicts the target value based on the nearest neighbors in the feature space. It calculates the average or majority vote of the target values of the k nearest neighbors. KNN can be effective when there is a correlation between the proximity of instances in the feature space and their similarity in terms of price. We can choose an appropriate value for k and evaluate the model using suitable metrics.
4) **Support Vector Regression (SVR):** SVR is a variant of support vector machines (SVM) that can be used for regression tasks. It finds a hyperplane that maximally fits within a specified margin of error (epsilon-tube) around the target values. SVR can capture non-linear relationships using kernel functions. It is important to tune the hyperparameters such as the choice of the kernel and regularization parameters for optimal performance.
5) **Random Forest:** Random Forest is an ensemble model that combines multiple decision trees to make predictions. It can handle non-linear relationships, interactions between features, and provide feature importance rankings. Random Forest can be effective in capturing complex patterns in the data. We can evaluate the model using appropriate regression metrics and analyze the importance of individual features.

When comparing these models, it's important to consider their performance on our specific dataset. We can use techniques such as cross-validation or train-test splits to assess their performance and compare the evaluation metrics. Additionally, we can analyze the interpretability of the models, computational requirements, and any specific assumptions they make to make an informed decision about the most suitable model for our mobile price prediction task.

## 4) Model Evaluation & Future Works

When evaluating the effectiveness of different regression models using the R2 score, it is important to understand the underlying principles and characteristics of each model. The R2 score measures the proportion of the variance in the target variable that can be explained by the independent variables. A higher R2 score indicates a better fit of the model to the data.

**Random Forest Regression:**

This model achieves a high R2 score of 0.932, indicating a strong ability to explain the variance in the target variable.

**Linear Regression:**

A moderate R2 score of 0.795 suggests that the linear relationship captures a significant portion of the variance in the target variable.

**KNN Regression:**

With an R2 score of 0.770, it indicates reasonable performance in capturing the underlying patterns in the data.

**Logistic Regression:**

The R2 score of 0.586 suggests that the model has limited ability to explain the variance in the target variable, indicating that logistic regression may not be suitable for this regression task.

**SVR Regression:**

The SVR model we evaluated has a negative R2 score of -0.178, which indicates that it performs worse than a simple mean prediction. This could be due to inappropriate hyperparameter tuning or a poor choice of the kernel function, leading to a weak fit to the data.

In summary, the models that work effectively, such as Random Forest, Linear Regression, and KNN Regression, demonstrate good performance because they are able to capture the underlying patterns and relationships in the data. On the other hand, Logistic Regression and SVR Regression perform poorly in this scenario, either due to their unsuitability for the regression task or their suboptimal configuration.

Some future works that could be added to our project are using other models like Deep Learning models & studying how to use it, using different datasets & comparing with this, trying to predict the prices of different regions at different time periods accurately. We could also try going into a specific brand in detail to try & get more accurate predictions.

## References

1) Mobile Price Class prediction using Machine Learning Techniques, March 2018

2) Mobile Price Prediction Using Machine Learning, July 22, 2020

3) Predicting Mobile Price range using Classification techniques, March 2021

4) Learn Mobile Price Prediction Through Four Classification Algorithms, 23 February 2022

5) Mobile Phone Price Prediction with Feature Reduction, February 2023

6) https://www.kaggle.com/datasets/rkiattisak/mobile-phone-price (Selected Dataset)

——————————————————————————————————————————