Mobile Phone Price Prediction Using Linear Regression and KNN Techniques
Data Analytics Assignment 2021-22
(TANISHQ VERMA 210642458)

***Abstract* – This paper contains finding and accurate predictions of Mobile Phone prices according to the different features provided to the model. Different types of supervised as well as unsupervised machine learning techniques have been applied to calculate the accuracy of the model on the selected dataset. The particularly selected dataset contains 21 dimensions and features which helps in correlating and predicting the approximate price of the mobile phone. This paper assesses the model and explains why the selected techniques have been applied to it and present the model's result. In this report, data preprocessing techniques have been applied to understand the dataset and visualize the dataset. We have given conclusions and result on assumptions and inferences based on the understanding of the model. A literature review has also been specified as part of all the references that have been taken while creating the model. The literature review will further help in discussions for the reader to help explore more domains. We have evaluated the dataset on Linear Regression, Knn Model and Logistic Regression Model to identify the pros and cons of the model and predict the accuracy for each of the specific models. Since in the market, the price of the mobile can vary depending on the features as mobile with the high specification can have a high price as well as low price depending on the maker of the device. The Machine Learning and Data Analytics techniques are used to give an approximate price of the mobile on the number of features the phone contains.**

***Keywords* – Logistic regression, Linear regression, Machine Learning, features**

I. INTRODUCTION

This project will accurately predict the price of the mobile phone using Linear Regression and Logistic Regression models which will both predict the price of the phone depending on the different features of the mobile phone are given to the model. This problem will help in understanding the customers the better way of what can be expected price of the phone based on what features they want. The customer can get a general range of an approximate price of the phone that they can expect. The dataset is available on Kaggle. The dataset was already having most of the features for us to use in our model and the features were relevant and related to computing the price of the phone. Further, the dataset had enough sample set to be used in our model.

As we have explored while creating the model, generally some mobile phones features can have a major impact on the price of the phone as compared to other features. To illustrate, the price of the phone can depend more on the release date, processor, storage as compared to whether the phone has Bluetooth, WIFI or color as these features are generally present in all kinds of phones. Usually, for the other features such as Bluetooth or WIFI, they tend to be present in almost all devices. As a result, these features become secondary as the customer going to buy a phone expect to get these features on the phone whether it is a high spec phone or a budget phone. In addition, while using it in our model, the data samples had contained these features in almost all the phones which further made it a less valuable feature.

Data Sample Mobile Phones

battery	color	clock	sim	FC	four_g	memory	m_dep	weight	cores
842	0	2.2	0	1	0	7	0.6	188	2
1021	1	0.5	1	0	1	53	0.7	136	3
563	1	0.5	1	2	1	41	0.9	145	5
615	1	2.5	0	0	0	10	0.8	131	6
1821	1	1.2	0	13	1	44	0.6	141	2

A. OBJECTIVES

- Identify the main features that have a significant contribution towards predicting the price of the phone.
- Select the different machine learning models that are best suited for our problem statement.
- Conducting preprocessing techniques to our data set to make an efficient model.
- Applying different data visualizing techniques including different scatter plots, box plots, point plots, etcetera to understand the dataset.
- Predicting the accuracy of the dataset on different machine learning models and understanding which is the most efficient amongst them.

II. LITERATURE REVIEW

This section of the report contains the different types of reports, online discussion, journals, literature reviews and references. The outcome and the assumptions for the same have been explained below.

The first report is titled 'Prediction of Phone Price Using Machine Learning Techniques' by Subhiksha S, Swathi Thota and J. Sangeetha [1] published under the SASTRA University. The aim of the paper was to calculate the price of the mobile phone using historical data that acts as the key the main aim to predict the approximate price considering the different types of features which contribute main and significantly towards the price factor. The report took different features such as Display, Processor, Memory, Camera and many into considering and calculating the price of the device. The report further discussed the different types of machine learning techniques

that can be used to create the model. Some of the different types of techniques were discussed in this project however the main focused techniques consist of Random Forest Classifier, Support Vector Machine and Logistic Regression. Based on the model generated by creating these techniques accuracy was predicted. The main focus of the report was to help the customers identify the right mobile with the best price. The report further explains that amongst the main three techniques used in the model, Support Vector Machine and Logistic Regression was able to achieve the highest accuracy of 81 percent.

The second report that was taken part literature review is titled 'Mobile Price Prediction Using Weka' by Pritish Arora, Sudhanshu Srivastava and Bindu Garg [2] by Bharati Vidyapeeth College of Engineering, Pune, India. The focus of this research work was to determine whether the given features could accurately predict the price of the mobile phone within a given range. The work consists of some specific selection algorithms which have contributing factors towards computing the price of the mobile phone and identifying and deleting features that are less necessary and redundant. The research was primarily focused to achieve maximum accuracy and choosing the minimum features. The conclusion was made based on the algorithm for the best selection of the features set and the best classification techniques for the given dataset. Some of the techniques that were used in this report include Decision Tree, Naïve Bayes and Machine Learning Techniques. The report was able to achieve the highest accuracy of 81.25 percent.

The final report for literature review is titled 'Predicting the price range of mobile phones using machine learning techniques' by K. S. Kalaivani, N. Priyadharshani, S. Nivedhanshri and R. Nanddhini [3] in AIP Conference Proceedings. This research work is recent and was published on 1 November 2021. Like the other two reports and research work, this research has a similar focus to calculate the price range of the mobile using the different types of features. Some of the features that were selected consist of memory, display, battery, camera, and some. The work is further used to create the model using the three main machine learning algorithms namely Support Vector Machine, Random Forest Classifier (RFC) and Logistic Regression. According to the report, they further reduced their features to 10 features only based on the contributing factor to the price range. After applying the data preprocessing techniques to their dataset, they applied features to the model consisting of main ML techniques. The report applies all the features to the model computed the accuracy for all three techniques. They further reduced their features to 10 and were able to increase their accuracy further. The final maximum accuracy that the model achieved was for SVM with 97 percent.

III. DATA MANAGEMENT

A. Data source and description

For this project, this dataset was present from the open-source website Kaggle and the link for it has been provided below.

The dataset consists of 20 features including one column for the index total comprising of 21 features. The data samples that were present are 2000. The shape for the data set is 21 x 2000. The dataset comprises different types of features such as battery, clock speed, front camera, RAM, height, width, memory and some other features as well total comprising of 20 features. Most of the features have the datatype of int expect clock and mobile depth which are both float types. The predicted value will be Price which is int value as well. According to the dataset, the data consist of a device that consists of different types of features such as colour, clock speed, RAM, Rear Camera and other but the most important feature of them all is the Price which is to be predicted by the model.

The information for the dataset has been given below

Dataset Features Type Information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 21 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   battery       2000 non-null   int64
 1   color         2000 non-null   int64
 2   clock         2000 non-null   float64
 3   sim           2000 non-null   int64
 4   FC            2000 non-null   int64
 5   four_g        2000 non-null   int64
 6   memory        2000 non-null   int64
 7   m_dep         2000 non-null   float64
 8   weight        2000 non-null   int64
 9   cores         2000 non-null   int64
 10  pc            2000 non-null   int64
 11  height        2000 non-null   int64
 12  width         2000 non-null   int64
 13  RAM           2000 non-null   int64
 14  sc_h          2000 non-null   int64
 15  sc_w          2000 non-null   int64
 16  talk_time     2000 non-null   int64
 17  3G            2000 non-null   int64
 18  touch_screen  2000 non-null   int64
 19  WIFI          2000 non-null   int64
 20  Price         2000 non-null   int64
dtypes: float64(2), int64(19)
memory usage: 328.2 KB
```

1) Data Preprocessing: For data preprocessing, we first checked for the missing values if present any in the dataset, we used the Pandas library where we used 'dropna' to drop any of the data samples that had any of the missing values. We further check it by using the 'shape' function which remained

unchanged implying that the dataset has no null values. For further preprocessing we used 'duplicate' to remove any of the duplicate data samples present in the data set. Finally, we used 'nunique' to further check for the unique values in the dataset. After applying both 'duplicate' and 'nuinque' we used 'shape' to cross verify the statement implemented above. Since all of the features are either integer and float values which will be used to compute the accuracy of the model, we do not need to type convert any of the features which would have been the case if there were any string or float values present.

2) Feature Selections: As part of Feature Selection, we could have used all of the features that are present in the data set, but for the optimal model, it is best to select the features that provide a significant factor to the predicted value which is the price in this case. Therefore, the feature that we selected are battery, color, clock speed, sim, RAM, touch and 3G. The features that we selected contribute and affect the pricing of the device as compared to any other feature and therefore the rest of the feature set can be ignored.

3) External Libraries: For this project, some of the libraries that we used are given below along with some of the descriptions of them.

NumPy – It is a library that helps and support the very large multi-dimension array and matrices. NumPy is used to provide access to large math libraries and solve difficult math functions for operating on these arrays and matrices.

Pandas – Pandas is a very popular library when it comes to Machine Learning models. It is a framework library in Python that can help to tackle data. A df or DataFrame is defined in 2 Dimension matrix that supports many of the operations on csv file and helps computations easy.

Scikit-learn – A Machine Learning library used in Python that is usually used in Data Science and Artificial Intelligence. It is a very straightforward and easy to use library used for implementing algorithms.

Mathplotlib- A library that is used for visualization purposes that consist of different types of graphs and plots. The legend and axis along with scales can be customized according to the need along with some changes in the color as well.

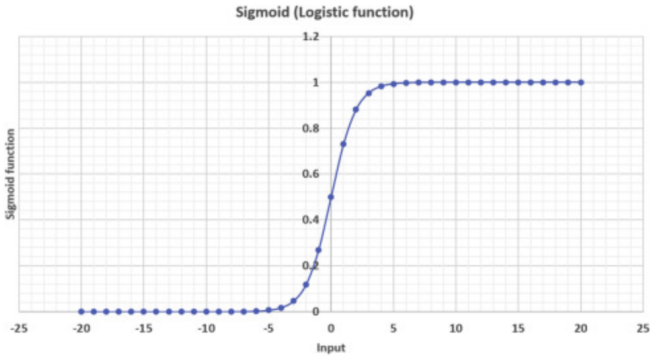
IV. METHODOLOGY

For creating the price of the phone, we use two models namely the Logistic Regression Algorithm and Linear Regression Algorithm. There are many pros and cons related to both the models but according to the dataset, we believe these are the best two algorithms for the model.

A. Logistic Regression Algorithm

Logistic Regression Algorithm is an algorithm that is used for optimization method based on boundaries of the classifier method. The range for the boundary is between 0-1 where weight and vectors have multiplied the result which will be corresponding to the distance from the boundary. For such an algorithm, the output if it is smaller or larger than the boundary help in the further prediction of which part of the instance of the predicted space is and if it further comes in the predicted space or not. To illustrate, whether the phone with these many features will belong to this price range or not. The P(x) is the value of the given logistic regression which predicts the probability of the instance belonging to the predicted domain and the 1-P(x) is the same instance but belonging to the opposite category of the predicted space.

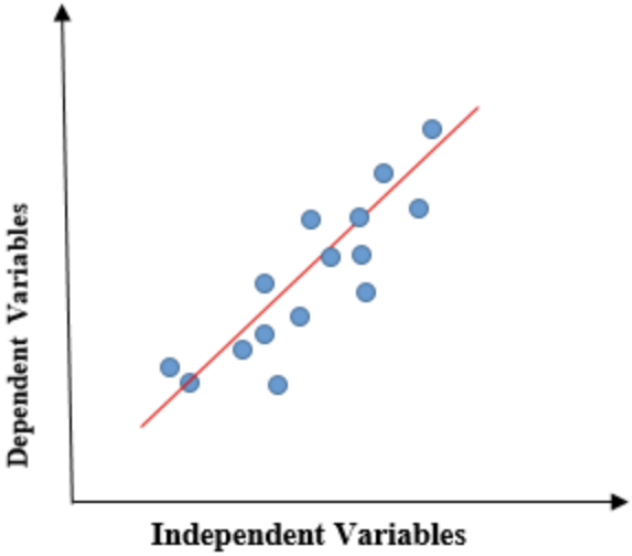
Logistic Regression Algorithm



B. Linear Regression Algorithm

Linear regression is simple and quite easy to understand method which is used for predicting analysis and showing a linear relationship with continuous variables. As the name suggests, Linear Regression shows a linear relationship with the independent variable of the x-axis and the dependent variables of the other axis. It gives a sloped straight line similar to that of Y = X depicting the relationship between the variables of both axes.

Linear Regression Algorithm



V. EXPLORATORY DATA ANALYSIS

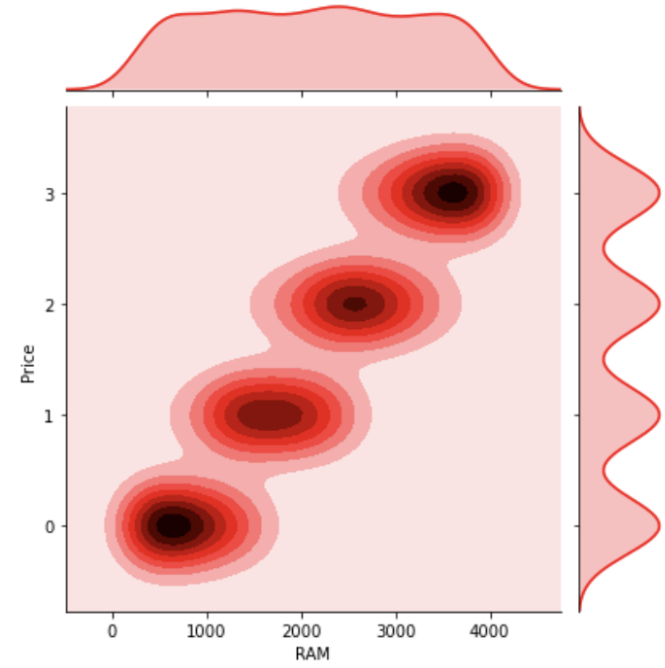
In the above section, we have understood the dataset and the feature selection which comprises the most important features that affect the price of the mobile phone significantly. A better way to understand the selected feature with different features is by visualizing the features with different features including the price of the mobile phone using the matplot library. As explained in the above section, the matplot library consists of various graphs such as boxplot, pointplot and others which will help to better understand the feature set with the price of the mobile.

According to the dataset mentioned in the feature selection part, since the RAM, battery and price both depend on each other significantly, it is best to visualize and see how each of them depend on each other vastly.

A. Price vs RAM

When we visualised the price of the device with each device's RAM, we can see from the jointplot that for each different price range of the device a range of RAM is present. This illustrates that a device present in one range of price can have multiple RAM. In conclusion, a device with a high price can have a low RAM or a device with a low price can have high RAM or vice-versa. However, one thing that jointplot visualisation depicts is that mobiles with a low price range have a smaller range of RAM as well.

Price vs RAM (jointplot)

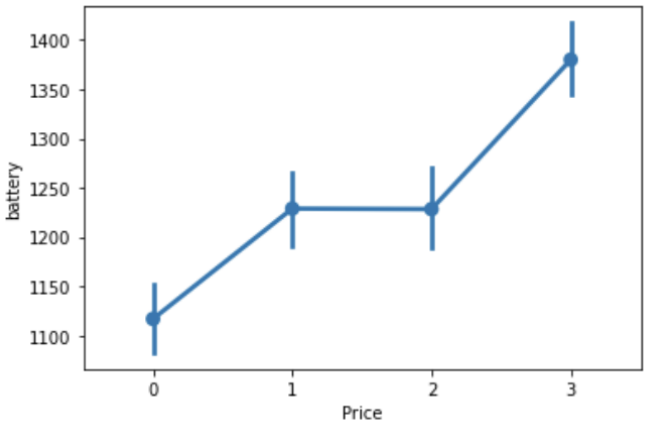


From the plot, we can see that the price range of the mobiles for 0 has a RAM range from 0-1000 and both has a direct correlation which each other meaning that as the range of the price increase so the RAM range increases as well.

B. Price vs Battery

When we visualised the price of the device with each devices Battery capacity, we can see from the pointplot that for each different price range of the device a range of battery capacity is present. The devices present in one range can have multiple battery capacities. Similar to the above relation, a device with a high price range can have a low battery capacity and a phone with a low price range can have high battery capacity. But, one thing that pointplot visualisation depicts is that mobiles with low price range have smaller range of battery capacity as well.

Price vs Battery (pointplot)

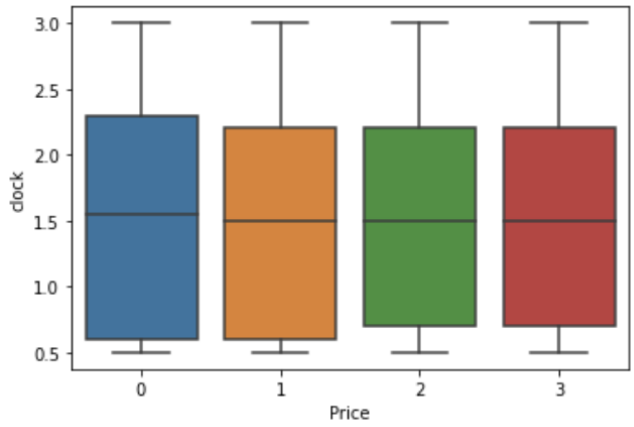


From the plot, we can see that the price range of the mobile for 0 has a Battery capacity from 800-1150 and both have direct correlation with each other meaning as the range of the price increase so the Battery capacity increases as well.

C. Price vs Clock Speed

When we visualised the price of the device with each device's Clock Speed, we can see from the boxplot that for each different price range of the device a similar range of Clock Speed is present. The device present is one range or be it in any other range it still has the same clock speed range. This concludes that a device with high price can have low clock speed or a device with a low price can have a high clock speed or vice-versa.

Price vs Clock Speed (boxplot)



From the plot, we can see that the price range of the mobile for 0 has a clock speed between 0.6-2.3 and both have a similar correlation with each other or it we can conclude that the price of the mobile cannot justify whether it will have a high clock speed or low clock speed.

VI. TESTING AND RESULTS

After complete visualization of different features with different plotlib and seaborn library, we were able to visualize and understand the relationship between different features. Following the visualization, we implemented the model to the different machine learning techniques however, we need to split the data into training and test so that we can predict the accuracy for each of the model or machine learning techniques and help understand which techniques are the best.

Splitting Data

```
a=data[['battery','color','clock','sim','RAM','touch_screen','3G']]
b=data['Price']
```

```
X_train, X_test, y_train, y_test = train_test_split(a, b, test_size
```

We split the data into 80-20 ratio meaning 80 per cent for training data and the remaining for testing data. We then further apply the techniques to these training and testing data and compute the accuracy for each of the techniques.

A. Linear Regression Model

After splitting the data, we applied training data to fit the model. Once our Linear Regression Model is fitted, we can compute the model with testing data using the score and check the accuracy of the model. For this dataset, we were able to achieve 87 per cent accuracy.

Linear Regression Model Accuracy

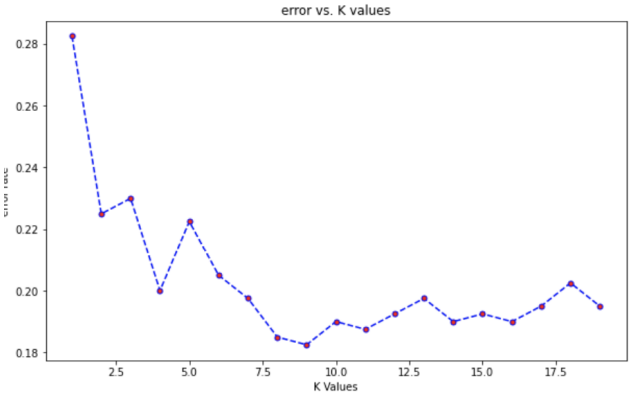
```
lrm.score(X_test,y_test)*100
```

```
87.11897187595284
```

B. K Nearest Neighbor Model

After computing the accuracy of the dataset on Linear Regression we further computer the accuracy of the dataset KNN model. However, we need to find the optimal value of K.

Elbow Method KNN



For this we applied the Elbow Method to compute the value of the error rate for different values of K. We took the value of K equal to 10 and then computed the accuracy of the model on the given dataset. We were able to achieve the accuracy of the dataset of 81 per cent.

KNN Model Accuracy

```
knn.score(X_test,y_test)*100
```

```
81.0
```

C. Logistic Regression Model

After computing the accuracy of the dataset on both above models, we finally applied our train and test data to Logistic Regression Model.

Logistic Regression Model Accuracy

```
lr.score(X_test,y_test)*100
```

```
59.75
```

For this method, we were able to achieve an accuracy of approximately 60 per cent which is far less than the other two models. Therefore, the above methods can be added to the model to get the optimal price of the phone based on the feature and this model can be ignored as the accuracy of this particular model is far less than the other two models.

VII. CONCLUSIONS

In this report, we computed the accuracy of the dataset on Linear Regression and Logistic Regression. We got the accuracy for both the models to be 87 per cent and 60 per cent respectively. From the accuracy score, we can estimate that for this data set Linear Regression is a better method as compared to Logistic Regression. Therefore, from the two models, Linear Regression is a better model.

Confusion Matrix

```
matrix=confusion_matrix(y_test,pred)
print(matrix)
```

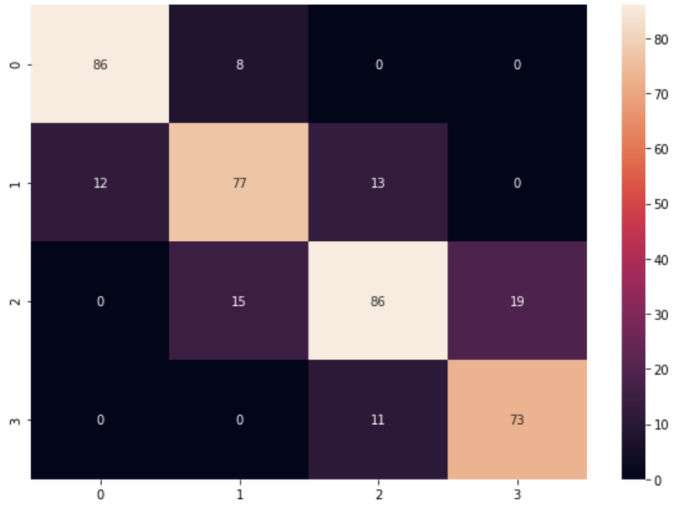
```
[[86  8  0  0]
 [12 77 13  0]
 [ 0 15 86 19]
 [ 0  0 11 73]]
```

When we applied the data to KNN model, we got an accuracy of 81 per cent which is a good score. Since KNN is a classifier model, we can check the correct and the incorrect prediction on the dataset by plotting the confusion matrix.

Heat Map

```
plt.figure(figsize = (10,7))
sns.heatmap(matrix,annot=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fdd801931c0>
```



Further, we can also plot a heatmap of the KNN model to check the false positive and false negative as well as the true positive and true negative for different categories of the data sample which can further help in explaining that this classifier model is working better on what categories of the data samples and is not able to make an accurate prediction on other specific categories of the data samples.

REFERENCES

[1] Prediction of Phone Prices Using Machine Learning Techniques Subhiksha S., Swathi Thota ,J.Sangeetha School Of Computing, SASTRA Deemed University Tirumalaisamudram,Thanjavur ,Tamil Nadu ,India - 613 401
URL: https://www.researchgate.net/publication/338471736_Prediction_of_Phone_Prices_Using_Machine_Learning_Techniques

[2] MOBILE PRICE PREDICTION USING WEKA Pritish Arora, Sudhanshu Srivastava, Bindu Garg Department of Computer Engineering, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune India
URL: https://www.ijsdr.org/papers/IJSDR2004057.pdf

[3] Predicting the price range of mobile phones using machine learning techniques K. S. Kalaivani, N. Priyadharshini, S. Nivedhashri, and R. Nandhini AIP Conference Proceedings 2387, 140010 (2021)
URL: https://aip.scitation.org/doi/abs/10.1063/5.0068605