ECS784P - Data Analytics - 2021/22
(Data Analysis on Mobile Price Prediction)

Abstract – **This paper contains findings of predictions made to predict mobile price prediction. In this project based on the mobile specifications like Battery, Bluetooth, Dual Sim, WIFI, RAM etc we are predicting the price range of the mobile. We are going to explore and analyse a data set which contains specifications of two thousand mobile phones, and we try to predict optimum price ranges for a list of mobile phones in the market by applying various machine learning algorithms such as Logistic Regression, k-nearest neighbors(knn) and Linear Regression.We have 2000 samples and 21 attributes. We assess the model and train it and explain why they were specifically used and then present the final prediction (Result). In the report it is also mentioned that how the data cleansing and analysis of the data have been done with proper understanding and examination. Here we are also showing the process of reaching conclusions based on the analysis. Literature review of the related work is also present in this report to expand the scope of talk for the reader**

Keywords – **(Logistic Regression, Linear Regression, k-NN, Machine Learning, Literature Review, analysis)**

I. INTRODUCTION

Price is the most effective attribute of marketing and business. Very first question of costumer is about the price of items. All the costumers are worried and thinks 'If he would be able to purchase something with given specifications or not'. We have chosen the mobile price prediction dataset which will help us to analyse all the price of the mobiles in the market for different specifications and requirements. The price of a product is the most important attribute of marketing the product. Now a days price maters a lot in a smart phone because it comes with a lot of features and company analyse it with full effort to price the phones which justifies the features included in the phone and covers the marketing and manufacturing costs of the mobile. And, mobile now a days is one of the most selling and purchasing devices. It has been observed that thousands of mobiles are purchased in a daily basis. Many features are important to be considered to estimate the price of mobile. Examples like processor, battery capacity which plays a major role because of the busy schedule pf the human beings. And from internal memory to camera everything gets consider with its top quality. And from browsing to texting and calling is one of the most important constraints in the technological era of 21st century

Here are the important attributes of our dataset: - Battery: Total energy a battery can store. - Bluetooth: Whether it has Bluetooth or not - 4G: Whether it has 4G or not - WiFi: Whether it has WiFi or not - RAM: Random Access Memory - Price Range: This is the target variable with value 0 (low cost), 1 (medium cost), 2 (high cost and 3 (very high cost))

We have 2000 samples and 21 attributes. The last attribute is a target attribute, which means that we have labelled data.

Here are the important attributes of our data set:
- Battery: Total energy a battery can store.
- Bluetooth: Whether it has Bluetooth or not
- 4G: Whether it has 4G or not
- WiFi: Whether it has WiFi or not
- RAM: Random Access Memory
- Price Range: This is the target variable with value 0 (low cost), 1 (medium cost), 2 (high cost and 3 (very high cost))

We have 2000 samples and 21 attributes. The last attribute is a target attribute, which means that we have labelled data.

Data Sample from original data set

Pixel Height	Pixel Width	RAM	Screen Height	Screen Width	Talk Time	3G	Touch Screen	WiFi	Price Range
20	756	2549	9	7	19	0	0	1	1
905	1988	2631	17	3	7	1	1	0	2
1263	1716	2603	11	2	9	1	1	0	2
1216	1786	2769	16	8	11	1	0	0	2
1208	1212	1411	8	2	15	1	1	0	1

A. OBJECTIVES

- To explore and analyse a datset which contains specifications of two thousand mobile phone
- Try to predict optimum price ranges for a list of mobile phones in the market.
- Highlighting possible improvements for further analysis
- To analyse the strength and weakness of models and evaluate all the aspects of it.
- Examining the exploratory data analysis to find the correlation between certain features in the dataset.
- Predicting the price of the mobile with respect to different features given in the dataset

II. LITERATURE REVIEW

In this part of the report literature review will be presented and dissected. The work dissected will be evaluated in relation to research problem being investigated.

The first report to be discussed is titled 'Mobile Price Class prediction using Machine Learning techniques' by Muhammad Asim and Zafar Khan. The aim of the report was to discuss the prediction of mobile price using different

This footnote will be used only by the Editor and Associate Editors. The edition in this area is not permitted to the authors. This footnote must not be removed while editing the manuscript.

machine learning algorithms which would further contribute to sales technology world for filtering the mobile phones with different specifications and requirements. According to conclusion of this paper the best feature selection algorithm and best classifies for the dataset used. This type of work can be used in any type of business model to find optimal product with minimum cost and maximum features. According to the paper it is also suggested that to find more sophisticated solution to given problem and more accurate tool for price estimation.

The second report that we will be discussing as a part of the literature review is titled 'Predicting the price range of mobile phones using machine learning techniques' by K.S. Kalaivani , N.Priyadharshini and S.Nivedhashri. According to the research paper the model is being developed to predict the price range of the new mobile phones in the market. Three machine learning algorithms namely Support vector Machine (SVM), Random Forest Classifier (RFC). Logistic Regression are used to train the model and predict the output as low, medium, high, or very high. The same thing is done for our model as well and trained it with Low, medium, and high. Here the data used is taken from Kaggle platform. And according to the paper to improve the classification accuracy, chi squared based feature selection method is used. Among the 21 features available in the dataset only up to 10 top features available in the dataset. As per the research before applying feature selection the accuracy obtained using svm, rfc and Logistic regression is 95 percentage, 83 percentage and 76 percentage respectively. And out of all the techniques it is been observed that svm gave the superior performance when compared to other tow classifiers.

The third report that we will be discussing as a part of the literature review is titled 'Predicting of Phone Prices using Machine Learning Techniques' by S.Subhishka, Swati Thota and J.Sangeetha. According to the research paper it looks to solve the problem by taking the data having the important features of smartphone along with its cost and develop a model that will predict approximate price of smartphone with reasonable accuracy. The dataset is used for the purpose for considering 21 different parameters for predicting the phone. Random forest, classifier, support vector mach9ine and logistic regression have been used primarily. This not only help customers to decide the right phone to purchase, it also helps the customers decide what should be appropriate pricing of the features they offer.

III. DATA MANAGEMENT

A. Data Source and Description

The dataset was taken from Kaggle, and it is an open-source website. The dataset has 2000 samples and 21 attributes. The last attribute is a target attribute which means that we have labelled data. The shape of the dataset is (2000, 21).
With our data we can classify the price range of mobile phones, so we are going to train a classification model to classify he price range as:
These are the target variables given below.
- 0(low cost)
- 1(medium cost)
- 2(high cost)
- 3(very high cost)

Described Data Sample

```
main_dataset.describe()
```

Front Camera	4G	Internal Memory	Mobile Depth	Mobile Weight	Processors	...	Pixel Height
00.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	...	2000.000000
4.309500	0.521500	32.046500	0.501750	140.249000	4.520500	...	645.108000
4.341444	0.499662	18.145715	0.288416	35.399655	2.287837	...	443.780811
0.000000	0.000000	2.000000	0.100000	80.000000	1.000000	...	0.000000
1.000000	0.000000	16.000000	0.200000	109.000000	3.000000	...	282.750000
3.000000	1.000000	32.000000	0.500000	141.000000	4.000000	...	564.000000
7.000000	1.000000	48.000000	0.800000	170.000000	7.000000	...	947.250000
19.000000	1.000000	64.000000	1.000000	200.000000	8.000000	...	1960.000000

The important attributes that are getting used for the prediction of the mobile prices are given below from our dataset:

- Battery: Total energy a battery can store
- Bluetooth: Has Bluetooth or not
- Clock Speed: Speed at which microprocessor executes instructions.
- Front Camera: Front camera mega pixels
- 4G: Has 4G or not
- Internal Memory: Internal Memory in Gigabytes
- Mobile weight: Weight of mobile phone
- Mobile Depth: Mobile depth in cm
- Processors: Number of cores processors
- RAM: Random Access Memory in Megabytes
- Talk Time: longest time that a single battery charge will last
- WiFi: Has WiFi or not

Data Sample from original data set

```
main_dataset.columns
```

```
Index(['Battery', 'Bluetooth', 'Clock Speed', 'Dual Sim', 'Front Camera', '4G',
       'Internal Memory', 'Mobile Depth', 'Mobile Weight', 'Processors',
       'Primary Camera', 'Pixel Height', 'Pixel Width', 'RAM', 'Screen Height',
       'Screen Width', 'Talk Time', '3G', 'Touch Screen', 'WiFi',
       'Price Range'],
      dtype='object')
```

B. Dealing with Missing Data

Before training the model, we need to pre-process the data so that it won't affect the accuracy of the model with Nan values or empty column space. As seen missing data can have a significant effect on the conclusions that can be established from the dataset. There are few ways or techniques to handle the missing/pre-processing the data shown below.

Pre-proceesing Techniques used

```
#Return DataFrame with duplicate rows removed
main_dataset = main_dataset.drop_duplicates()

#Remove missing values.
main_dataset = main_dataset.dropna()
main_dataset.shape

#Return number of unique elements in the object
main_dataset.nunique()
main_dataset.shape
```

Note : We do not have any null values in our dataset. It will ease the pre-processing technique

C. Feature Selection

Our database contains 21 features but from the analysing perspective only few of them were relevant for machine learning.This database contains 76 features, but the Cleveland data source concluded that with professional knowledge only a subset of 14 of them are relevant for machine learning due to the remaining features being very redundant based on predicting the price pf the mobile. We have 2000 samples and 21 attributes. The last attributes are a target attribute which means that we have labelled data.

D. External Libraries

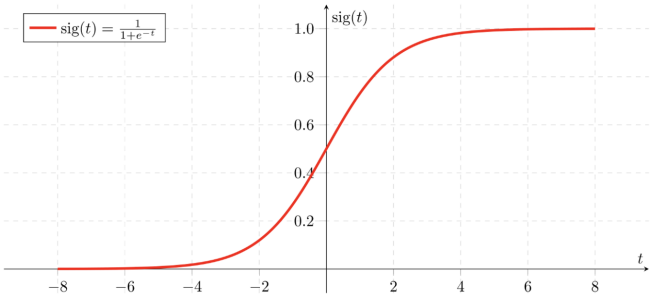
- NumPy: A library that includes support for large multidimensional arrays and matrices, extensively used as it provides access to very large library of advanced level math functions for operations on these arrays and matrices
- Pandas: A popular data framework library in Python one can manipulate and present data. A data frame is defined as 2D matrix that supports many operations on it for convenience.
- Scikit-learn: A machine learning library in Python that is extensively used within Data Science and Computer Science. It was used in this project for the implementation of the algorithms as it is easy and straightforward to implement.
- Matplotlib: A library for visualisation of data there are different types of graphs that can be developed using this library. The graph's axis and scales can be changes accordingly and some elements of customisation in colour are possible.

IV. METHODOLOGY

A. Logistic Regression

It is an optimization method that is based on boundaries in classifier models. A boundary is set usually between 0-1 when the weight and vectors are multiplied the outcome will correspond with the distance from the boundary. So basically, we used this method as a classifier but not as an optimizer. A P(X) value is given because of the logistic function determining the probability of an instance belonging to a section in the prediction space, 1-P(X) is the probability of that same instance belonging in the opposite categorical section.

Sigmoid Function



B. k-Nearest-Neighbor

The k-Nearest-neighbor (kNN) is a simple but effective method for classification. The major drawbacks with respect to kNN are:

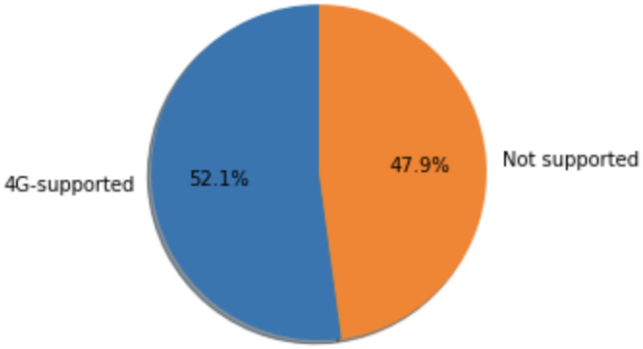
- Its low efficiency – being a lazy learning method prohibits it in many applications such as dynamic web mining for a large repo.
- Its dependency on the selection of a 'good value' for k

C. Linear Regression

Linear Regression plays a fundamental role in statistical modelling. This article provides a step-by-step coverage of linear models in order of model specification, model estimation, statistical inference, variable selection, model diagnosis and the prediction. It can be utilized to assess the strength for the relationship between variables and for modelling the future relationship between them. It includes several variations such as linear, multiple linear and nonlinear. The most common models are simple linear and multiple linear.

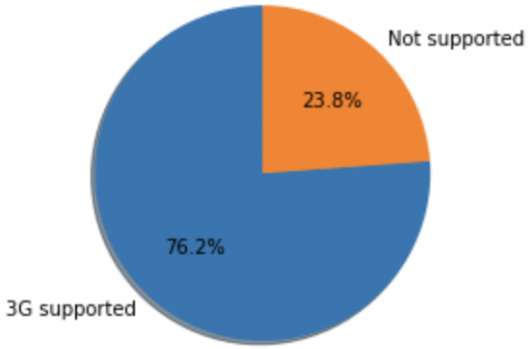
V. EXPLORATORY DATA ANALYSIS

As we are familiar with our dataset now and we know what attributes we have and which ones are important, we will perform some experiments and conduct data analysis with our dataset so that we can have a general idea of interrelationships. From our dataset we can easily analyse that what percentage of phone supports 4G and what percentage of phones don't support 4G .We will use pie-plot from matplotlib to visualise our findings and it will give us idea how much percentage of phones support 4G and don't support 4G. As from pie chart we found out that 52.1 percentage of the phones support 4G and rest 47.9 percentage don't support 4G. To continue our analysis, we now want to look at the distribution. From the distribution we have labelled '4G- supported' by blue and by orange for 'Not Supported' .



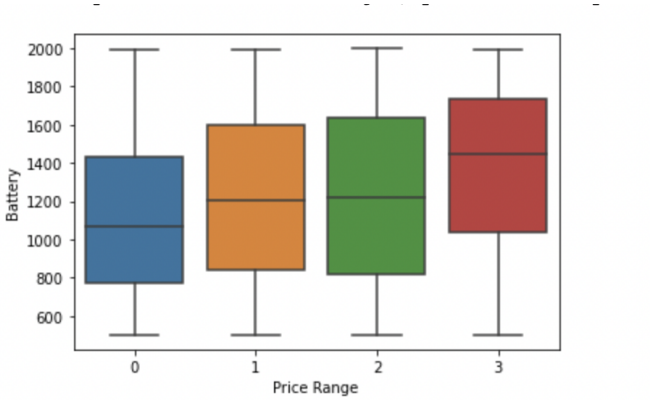
4G Data Visualisation

As we are familiar with our dataset now and we know what attributes we have and which ones are important, we will perform some experiments and conduct data analysis with our dataset so that we can have a general idea of interrelationships. From our dataset we can easily analyse that what percentage of phone supports 4G and what percentage of phones don't support 4G .We will use pie-plot from matplotlib to visualise our findings and it will give us idea how much percentage of phones support 4G and don't support 4G. As from pie chart we found out that 52.1 percentage of the phones support 4G and rest 47.9 percentage don't support 4G. To continue our analysis, we now want to look at the distribution. From the distribution we have labelled '4G- supported' by blue and by orange for 'Not Supported' .



3G Data Visualisation

For more analysation we have visualised the dataset to predict the price of the phone by battery capacity. The better and long running battery means the price of the phone will be higher as compared to battery of lower price.We will use box-plot from sns to visualise our findings and it will give us idea of the range of the mobile phones with respect to battery capacity.



Battery Capacity vs Price Range

To continue our analysis, we can look at the ranges distributed. For price range these are the target values shown below:

- 0(low cost)
- 1(medium cost)
- 2(high cost)
- 3(very high cost)

If we see the graph for price range 0(low cost) we can observe that the battery capacity lies between 800-1400 mAh and similarly for price range 3(very high cost) we can observe that the battery capacity lies between 1100-1700 mAh And from this we can conclude that mobile phones with higher battery capacity will be costlier.

VI. TESTING AND RESULTS

A. Splitting the data set

We have separated the dataset into training set and testing set, most of the data is used for training and the smaller portion of data is used for testing. Analysis services randomly samples the data to help ensure that the testing and training sets are similar and then we train the model using the training set and then apply the model to test set and with this we can evaluate the performance of our model. In our case we have split 80 percentage of data for training and 20 percentage of data for testing.

B. Training data with Linear Regression

With Scikit-Learn it is extremely straight forward to implement linear regression models, as all you really need to do is import the Linear Regression class, instantiate it, and call the fit() method along with our training data. This is about as simple as it gets when using a machine learning library to train on your data

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

```
91.05323761864229
```

Linear Regression Accuracy

C. Training data with Logistic Regression

For this technique we use the training and testing data to predict the accuracy/score of the model. We make an instance of the model then we fit the data using logisticRegr.fit(x_train, y_train)*and predict the accuracy using the test data.*

```
LogisticRegression()
```

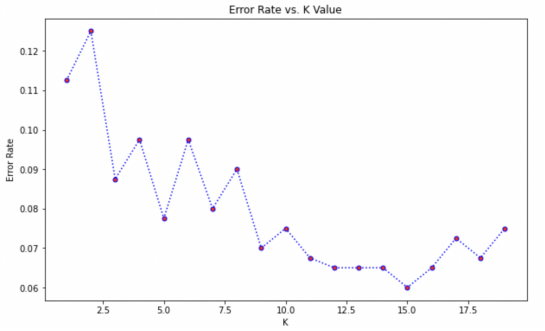
```
logmodel.score(X_test,y_test)*100
```

```
68.0
```

Logistic Regression Accuracy

D. Training data with k-NN

E. elbow method for optimal value of K



Error Rate vs. K Value

'K' is the number of nearest training points which we classify them using the majority vote. Here we calculate the value of K using elbow method, elbow method helps us to select the optimal number of clusters for KNN clustering.

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

```
92.5
```

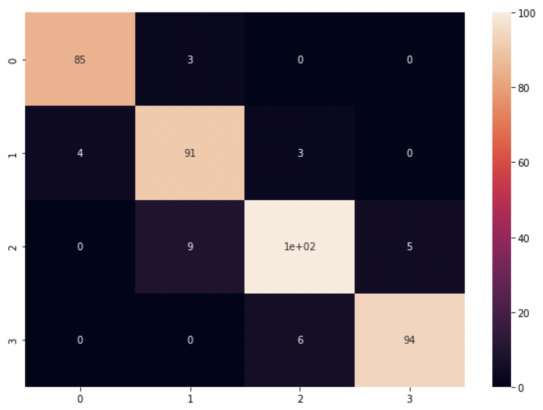
k-NN Accuracy

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

```
[[ 85   3   0   0]
 [  4  91   3   0]
 [  0   9 100   5]
 [  0   0   6  94]]
```

Confusion Matrix

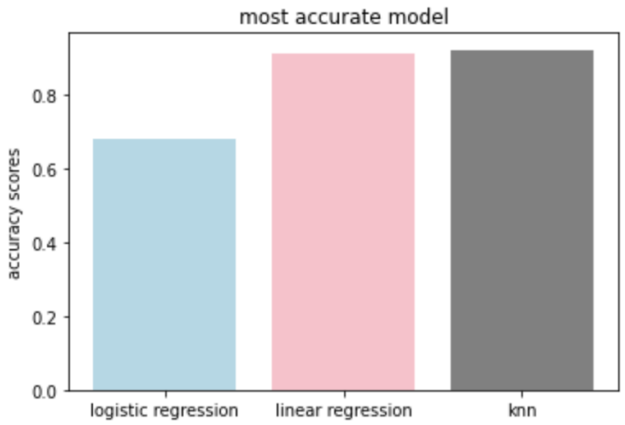
We are using confusion matrix to describe the performance of our classification model.



Heat Map

We use heatmap to represent various shades of the same colour for each value that is needed to be plotted. It is also observed that darker shade of chart always represents higher values.

VII. CONCLUSIONS



After training our dataset with 3 different models we can conclude that KNN is the best model for our dataset as it has the highest accuracy rate i.e., 92.5 percentage and now finally we can run our k-NN model to predict target values on the test dataset.

VIII. REFERENCES

[1]Mobile Price Class prediction using Machine Learning Techniques Muhammad Asim, Zafar Khan International Journal of Computer Applications URL https://www.researchgate.net/profile/Muhammad-Asim-9/publication/323994340_Mobile_Price_Class_prediction_using_Machine_Learning_Techniques/links/5b2b23b94585150c63446830/Mobile-Price-Class-prediction-using-Machine-Learning-pdf

[2]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 URL https://aip.scitation.org/doi/abs/10.1063/5.0068605

[3]Prediction of Phone Prices Using Machine Learning Techniques S. SubhikshaEmail authorSwathi ThotaJ. Sangeetha School of ComputingSASTRA Deemed UniversityThanjavurIndia URL https://link.springer.com/chapter/10.1007/978-981-15-1097-7_65