Mobile Price Prediction using Supervised and Unsupervised Learning Techniques

Ayush Sanghavi, Milind Vakharia, Vedant Benadikar, Vighnesh Kolhatkar

Department of Computer Science, Luddy School Of Informatics, Computing and Engineering
Indiana University Bloomington

ABSTRACT

Smartphone ownership has become extremely essential in today's world. The manufacturing of Mobile devices in 1980's was highly expensive. Back then, it was a luxury and now it has become a necessity. Using Machine Learning, we can gauge a lot of insights regarding the features and the price of a mobile device so the affordability could be scaled for every class. As we gather more and more data, we understand the usage of Mobile devices and its prices in a better way and hence can increase the accuracy of our models to give high efficient, modular and close to precise data.

1. INTRODUCTION

The project will help to analyze the price on the basis of a lot of features available today. Using models like SVM, Logistic Regression, Decision Tree, Random Forest, ANN, we classify the multi valued target feature; 'Price_Range'. We will visualize the data to give us a clearer perspective which will give us a perception that will help people choose the right mobile phone with efficient use. For example, some people may not require extremely good camera but a good battery life for the same price range and may end up buying the wrong mobile phone.

2. RELATED WORK

Most research papers we referred to, had implemented one Machine Learning model to predict the price or the price range. We have executed multiple supervised ML models, created a deep learning neural network using ANN, and K-means Clustering to cluster the price ranges of mobile phones depending on the features. We then determined our observations that analyze which model best fits our data thereby giving high accuracy.

Using previous data to predict the price of available and new launching products is an interesting research background for machine learning researchers.

The price of cars and mobiles can be compared, because the prices of both depend on the features given by the manufacturer.

Sameerchand-Pudaruth [1] predicted the prices of second hand cars in Mauritius. He implemented many techniques like Multiple linear regression, k-nearest neighbors(KNN), Decision Tree, and Naïve Bayes to predict the prices. Sameerchand-Pudaruth got Comparable results from all these techniques. During research it was found that most popular algorithms i.e Decision Tree and Naïve Bayes are unable to handle, classify and predict Numerical values. Number of instances for his research was only 97(47 Toyota+38 Nissan+12 Honda). Since he had less number of instances, the accuracies turned out to be low. [1].

Muhammad Asim, Zafar Khan [2], published about Mobile Price Class prediction using Machine Learning Techniques to predict If the mobile with given features will be Economical or Expensive, they collect data from the website www.GSMArena.com. When they used decision tree classifiers on 28 instances, their model correctly accurate 20 instances of them; i.e. around 71.42%. [2]

There has been quite a lot of research done in this field. The work done in this field has been inspirational and there have been various findings which we corroborate. Most of the research in this field consists of only predicting the mobile price using a singular method. These are predicting the price, but we do not know whether the method used is the best method to predict the target value. We also don't know much about the data itself. Hence we have decided to visualise and understand the data first, which will then make it easier to understand the problem and select the best features. We have also used various supervised and unsupervised learning models to classify the price range and compare them to know which model gives us the best accuracy and what is the reason behind it. We decided to train an artificial neural network using keras and implemented the model using different parameter settings to predict the price range with the highest accuracy.

3. APPROACH USED

Multiple approaches of classification and regression in the field of Supervised and Unsupervised Learning has been used. The following steps were performed to tackle this problem:

1. Loading the dataset, pre-processing, cleaning data and removing all unwanted data like missing values in the target attribute or negative classes

present in a column with positive classes. All columns were also renamed for better readability.

2. Visualizing our Data, finding correlation among features and target label (Price_Range). Finding the most important attributes in the dataset and their dependencies. It was observed that the attribute 'RAM' had the highest correlation with the target attribute. Another attribute called 'Battery_Power' had significant correlation with the target variable. Bagged Decision Trees like Extra Trees Classifier were used to confirm the above observations.

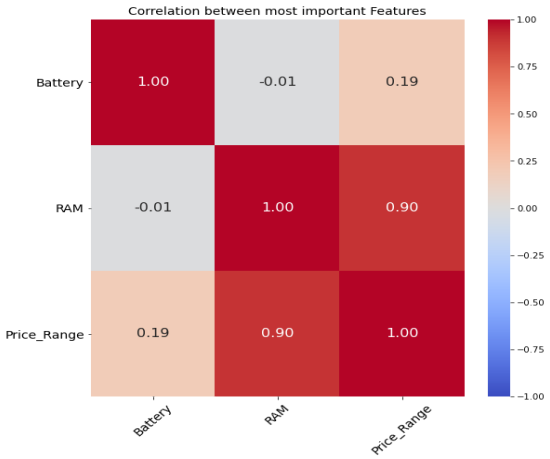

Fig 1. Threshold heatmap

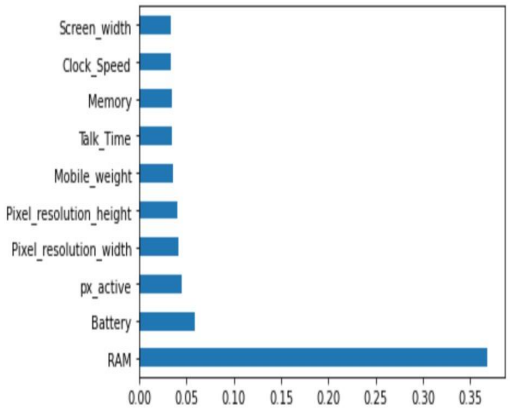

Fig 2. Extra Trees Output

3. Splitting the data into training samples and testing samples. Split used: 80%(training) - 20%(testing).

3.1 Classification using different approaches

Using multiple classification techniques using different classifiers and finding the accuracy of all the models. The different classifiers used on this dataset are Decision Trees, Random Forests, Support Vector Machines, Gradient Boosting, Logistic Regression and KNeighborsClassifier. The following confusion matrices were obtained:

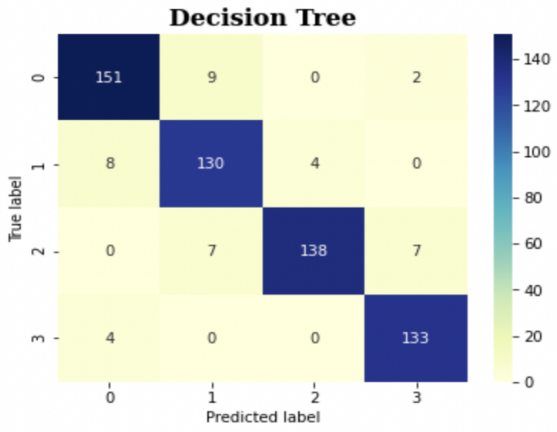

Fig 3. Confusion Matrix of Decision Tree

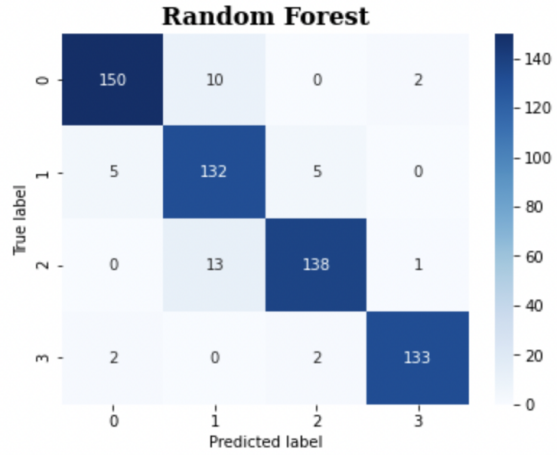

Fig 4. Confusion Matrix of Random Forest

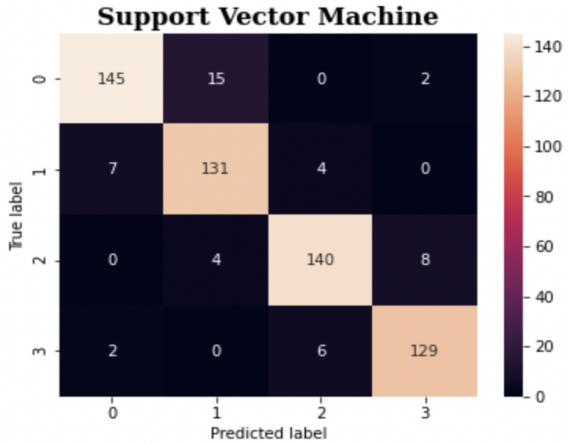

Fig 5. Confusion Matrix of Support Vector Machine

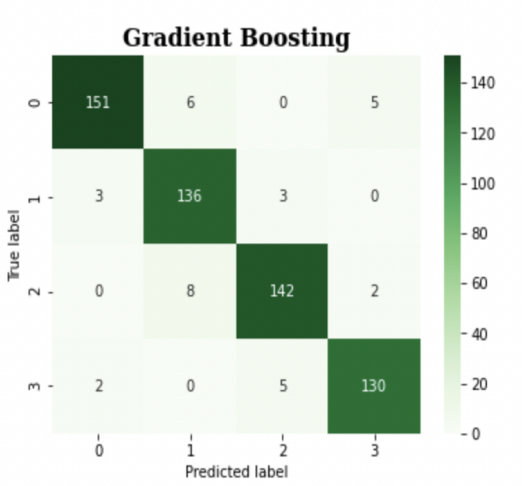

Fig 6. Confusion Matrix of Gradient Boosting

Fig 7. Confusion Matrix of Logistic Regression

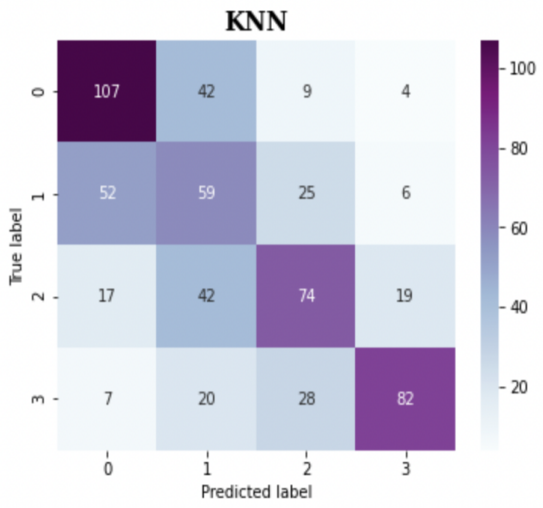

Fig 8. Confusion Matrix of KNN

Analyzing and concluding the best model for classification. It was observed that Random Forest classifier and Logistic Regression gave really good accuracy. Classification was performed on two more attributes: 3G and 4G. Similar trend for these columns for all classifiers can be observed. Due to high dimensionality, KNN did not give a good accuracy.

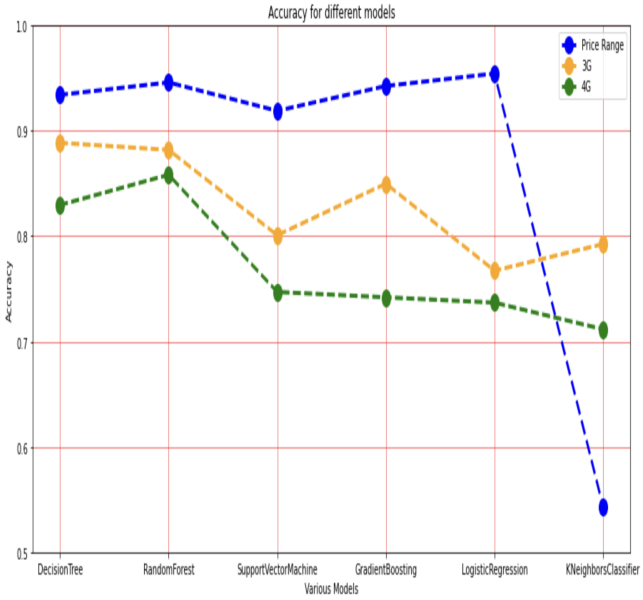

Fig 9. Accuracies of different classifier models

The accuracy for the target column is the highest across all classifiers apart from KNN. This is mainly because of the number of dimensions. PCA can be performed to reduce the number of dimensions, especially by removing the least important attributes. This should increase the accuracy for the KNN classifier as well.

3.2 Predicting target attribute using Deep Learning

We have used Tensorflow and keras library to successfully implement and train our artificial neural network model. We have used Tensorflow because it is a flexible ecosystem of tools, libraries and resources that lets developers easily build and deploy ML powered applications. The reason we used the keras library is because we noticed that it follows the best practices for reducing our cognitive load. It also offers consistent & simple APIs which minimizes the number of user actions required and makes it easier to use.

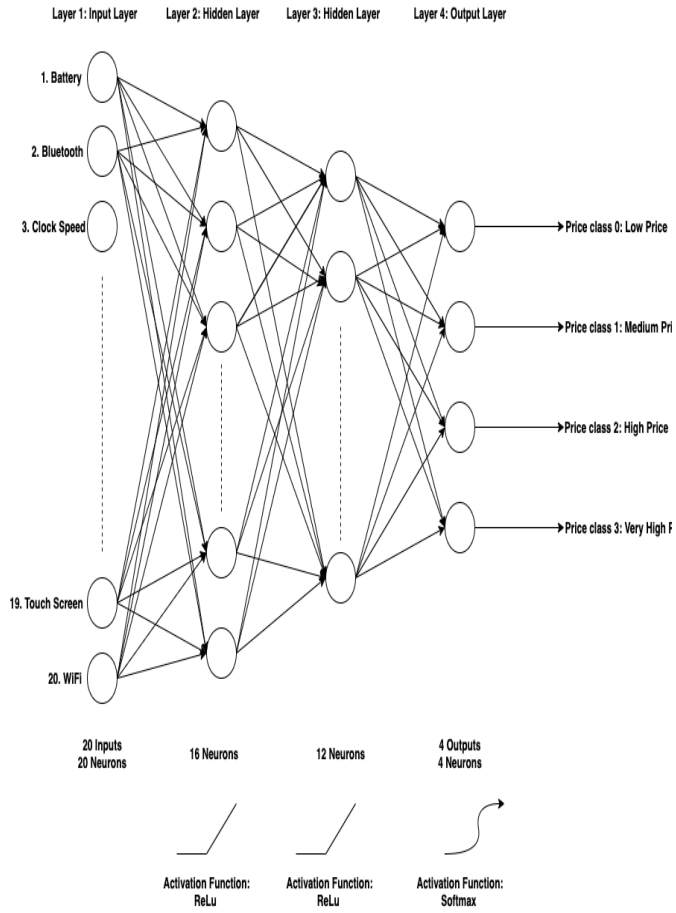

Fig 10. Neural Network and it's parameters

Neural networks involve input and output layers as well as hidden layers which will transform the input into something that the output layer can see.

- The input layer consists of the 20 different features which are present in our dataset ranging from Battery to Wifi.
- The output layer consists of the multiclass target attribute which is price range which can be 4 values 0,1,2,3 where 0 is the lowest price and 3 is the highest price.
- There are two hidden layers as shown in the figure which are of 16 and 12 dimensions respectively.
- We have used the ReLu activation function for the hidden layers, and Softmax activation function for the output layer since the output is a multiclass attribute.
- Adam optimizer is used to shape and mold the model into its most accurate possible form by changing the weights.
- Our loss function is categorical cross entropy as it is used when we have multiple classes as output.
- Finally, we used a batch size of 64 and kept the number of epochs as 100.

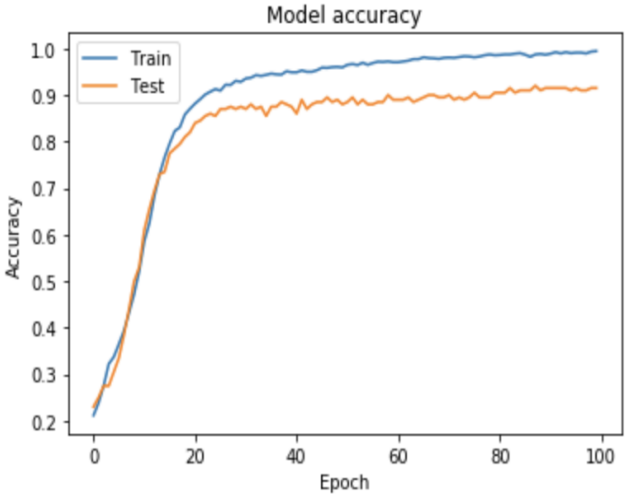

Fig 11. Model accuracy on training and testing data

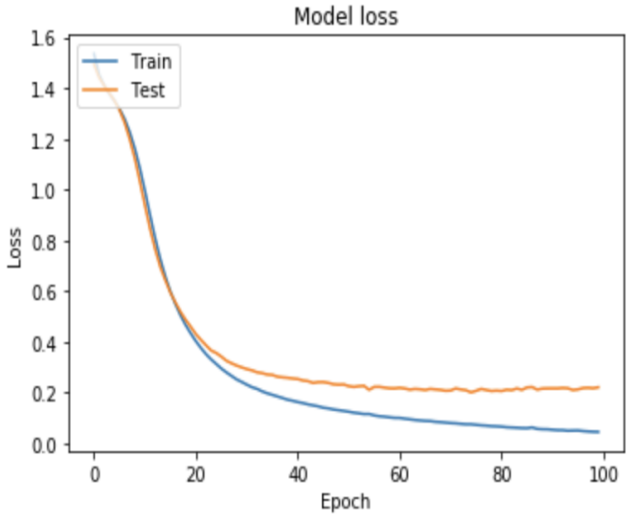

Fig 12. Model loss on training and testing data

As we can see in the above outputs, the ANN has been trained well and gives us a good accuracy (93%) to predict the target attribute (price range). Also, the loss keeps reducing as the number of epochs increase, which is what we have expected.

3.3 Performing unsupervised K-Means Clustering.

Multidimensional Scaling was implemented on the dataset. The K-means clustering model was fit on this multidimensional scaled data and the labels were plot using seaborn.

After comparing K-means labels with the ground truth, we could see that K-means could not group the classes well. The adjusted Random score indicates that predicted labels were very close to the random labels and it shows when

plotting the multidimensional scale data with the predicted labels.

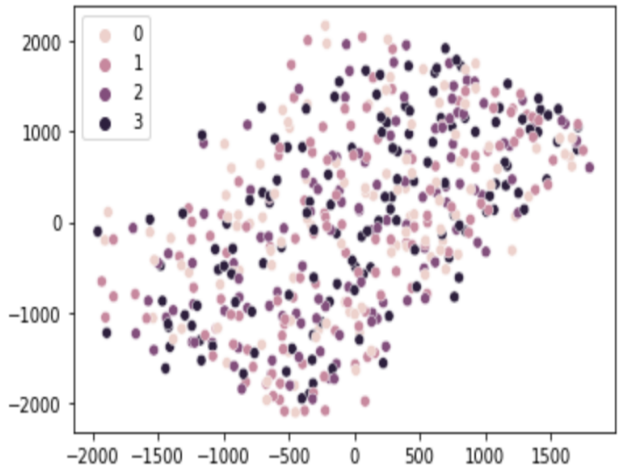

Fig 13. K Means Clustering

4. CHALLENGES AND LEARNING

10.1 Our models would perform much better if we had more data.

10.2 There were some features in the dataset that were outliers and were ambiguous. We had to drop the outliers, missing values and wrong data.

10.3 If we had more area specific data, we could add regional requirements for mobile devices which would help the masses over the globe.

10.4 We learnt a lot about different Machine Learning models and how it could be implemented in the real world.

10.5 We understood which model would best fit our data as well as given such data, which algorithm would output high accuracy.

5. FUTURE SCOPE

11.1 This project can be deployed as a web application, where we can help users understand the price range of their mobile device based on different features.

11.2 We can add additional features such as interactive dashboards where we can drag the features, design the desired device and classify or predict the price range of the mobile phone.

11.3 By adding more and more real world data, we can achieve maximum accuracy. This will help our models train in a way that helps us give maximum outcomes.

11.4 Using better AI techniques that will give a list of your preferred device in a priority order so you can choose correctly.

6. REFERENCES

[1] Sameerchand Pudaruth . “Predicting the Price of Used Cars using Machine Learning Techniques”, International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 7 (2014), pp. 753- 764

[2] Muhammad Asim and Zafar Khan. Mobile price class prediction using machine learning techniques. International Journal of Computer Applications, 179(29):6–11, 2018. doi: 10.5120/ijca2018916555.

[3] ANN for Predicting Mobile Phone Price Range: https://philarchive.org/archive/KHIAFP.

[4] Mobile Price Class prediction using Machine Learning Techniques
:https://www.ijcaonline.org/archives/volume179/number29/29158- 2018916555

[5] Mobile Phone Sales Forecast Based on Support Vector Machine::
https://iopscience.iop.org/article/10.1088/1742-6596/1229/1/012061/pdf