

M. Abdullah Tayyab (22-11210), Zohaib Amir (22-10462), Asim Arif (22-11211), Ahad Tariq(21-11102) , and Usama Sohail(21-10054)

CSCS455 – Section A – SP21

Data Mining Project

**Abstract:**

The main motive of this paper is to predict “the prices of mobile phones and if they are expensive or economical with the given features”. For this purpose, the data set is collected from Kaggle. To attain minimum computational complexity, different types of feature selection algorithms are used to remove minor and less important features and focus on the major ones. To engineer the highest accuracy possible, different classifiers are used for. Results are obtained on the basis of a comparison of the minimum features selected and the highest accuracy achieved. After that comes the part of conclusion and it is made on the basis of the best feature selection algorithm and best classifier used to predict the prices of mobile phones with the given features.

**Introduction:**

This paper focuses on predicting the price of mobile phones considering their features and then identifying if they are expensive or economical. We live in an era where mobile phones are the ones that are the most in demand because it has become a necessity in our daily life, from teenage years to adults, everyone owns a mobile phone. Whenever a person is supposed to buy something be it a mobile phone, a house, or a car, the major concerns in his head are about the price of the product, features being offered to them and wondering if it can be afforded or not or comes in the bracket of economical or expensive for him. The paper aims to help users to predict the most accurate price for a mobile phone to help the users in getting their desired phone offering maximum required featured according to what they can afford; helping a user get an optimal mobile phone for them.

Artificial Intelligence has come a long way and is helping the world in countless ways. There are many fields and many researchers who have tried to predict pricing of different products for instance, homes, cars etc. By using different tools, techniques, and algorithms like regression, supervised and unsupervised learning and classification etc., AI helps us to achieve our goal of predicting the price for a mobile phone.

Countless mobile phones are bought on the daily basis for numerous purposes considering the features a mobile phone provides to its consumers. Talking about features, there are many features that are taken into consideration these days that includes, for starters, the battery life, the quality of camera, 4G/5G supported devices, size of the screen, internal memory size, RAM size, the length, thickness, and weight of a mobile phone etc. These are few important features that are commonly considered while purchasing a mobile phone.

The following paper consists of research and analysis of related work done by other researchers. The following section consists of methodology, the algorithms and techniques used to achieve our desired goal, and the comparisons made to evaluate the results. The final section will focus on concluding the topic based on our research.

**Literature Review:**

Advancements in the artificial intelligence have helped in addressing and resolving many real-life problems using its different algorithms and tools. Predicting the price of a product for instance mobile phones, houses, cars, apparels etc., is one of the problems that is continuously being researched and worked. Many researchers have worked on predicting the prices of certain products using various different techniques like, k-nearest neighbors (KNN), linear regression, decision trees, multivariate regression, multiple regression etc., all these techniques and algorithms are being used in order to predict the price of a certain product.

There have been many researchers who have experimented to predict the price of products using various techniques like decision trees, multiple linear regression, KNN etc. One of these researchers are Kanwal Noor and Saddaqat Jan [1], they worked on providing predicted price of a vehicle by using multiple linear regression and the price prediction used various factors or features that included, alloy rims, steering type, vehicle type, etc. They were able to get maximum accuracy on these factors and the main technique and algorithm used by them was multiple linear regression.

Various research for predicting the prices of second-hand vehicles have also been made. Sameerchand Pudrauth [2] worked on this and used various techniques for this purpose that mainly include, decision trees, Naïve bayes, multiple regression and KNN. The research and experiment were not up to mark and it failed because of inconsistency. The reason that poor accuracy in the predicted price was found is because of the incapability of the techniques used Decision Trees and Naïve Bayes. These two algorithms are very common yet they are not capable to handle and classify numeric values, especially in a large data set.

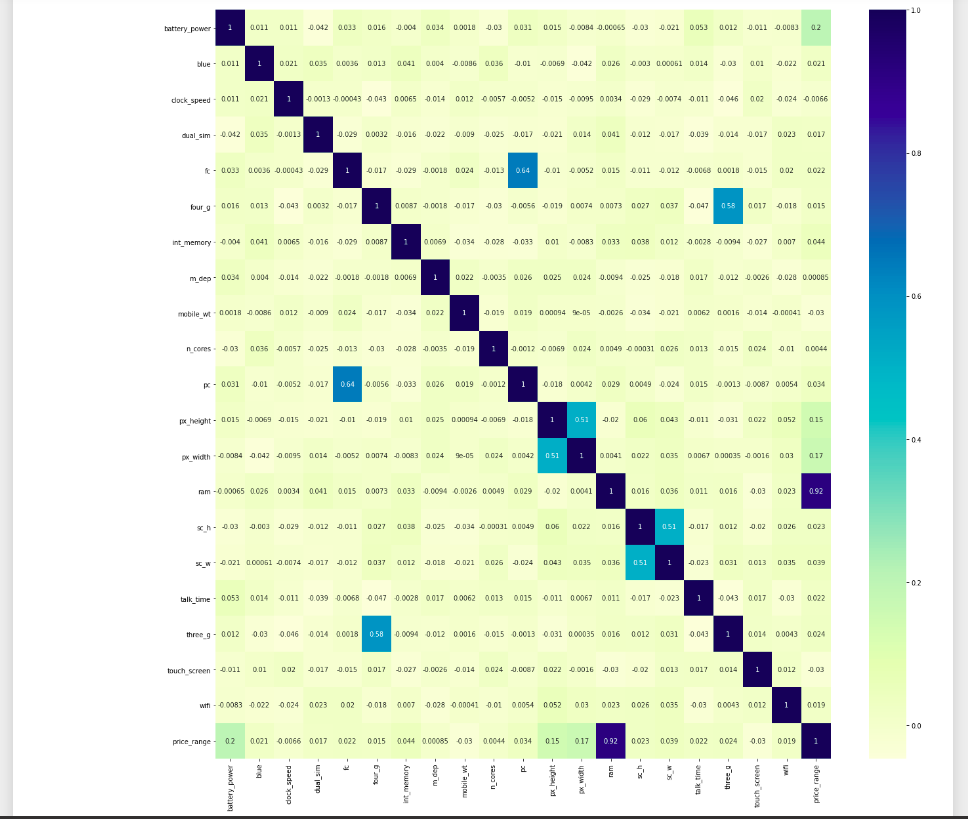
Another research done by Limsombunchai [3] was studied for this paper. Limsombunchai’s [3] research was on identifying the better methods used for giving estimated prices of a house. It was concluded in the research, that Neutral Networks was much better and efficient method than the Hedonic method because the former method is first trained and then evaluated unlike in the latter one.

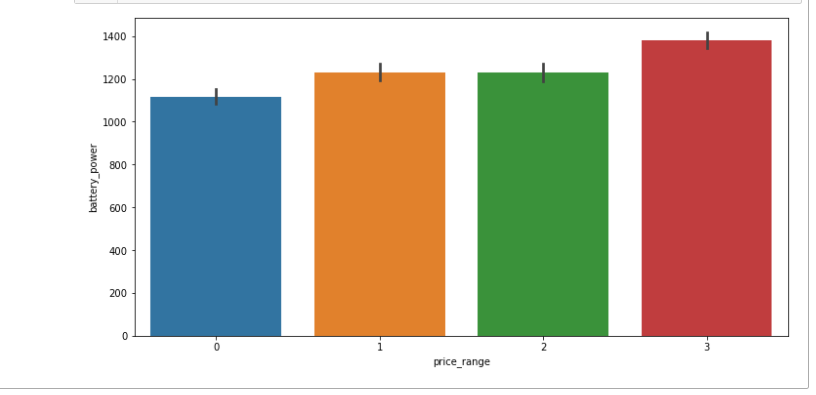
The paper *“Introduction to Multiple Regression: How Much Is Your Car Worth?”* [Shonda Kuiper, 2017], focused on providing students and researchers the better methods and algorithms for predicting the cost of products. The method adopted by Shonda Kuiper [4] was multivariate regression to predict the price of a used car. The paper is quite helpful as its study shows which methods should be avoided and which ones should be used to help in predicting the prices of products with similar attributes and data set.

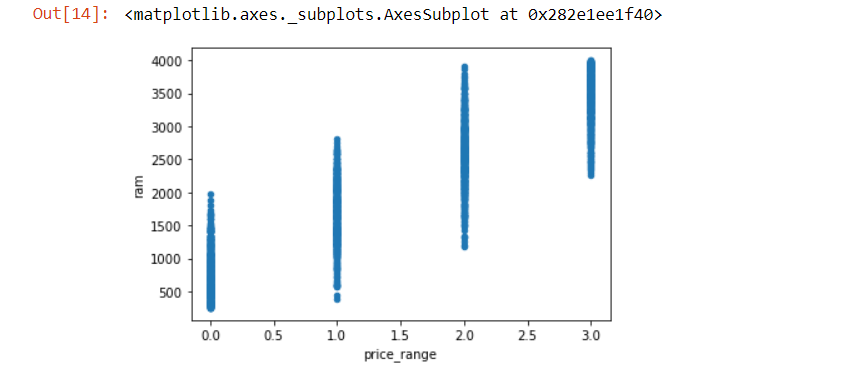
Studying different researches and experiments previously done by others helped us model our research paper that focuses on providing predicted prices of a mobile phone considering the features being offered. Algorithms, models, and techniques used by others, the accuracies and inaccuracies found in their paper gives us quite a useful information to make sure we use the correct machine learning techniques to achieve our desired results.

**Methodology:**

For the purpose of performing Mobile Price Classification five classifiers are being used and compared. The five classifiers are Decision Trees Classifier, Logistic Regression, XGBoost, Support Vector Machine, and K-Nearest Neighbors. We have used the data set from Kaggle called Mobile Price Classification. The data set consisted of 21 features in total with the target variable being the price. The target variable had 4 classes. The classes were low cost (0), medium cost (1), high cost (2), and very high cost (3). The features were battery power, front camera pixels, Bluetooth, ram, internal memory, and many more. In the beginning data pre processing has been performed. The data set was checked if it contained any missing values. After this, data visualization techniques using the library Sea born were performed in order to check the relationship between the features and the target variable. Most of the features had a weak correlation with the target variable except ram which had a very strong relationship with the price range. Battery power didn’t have a weak relationship but did a have a moderate relationship with the price range to an extent. A heat map, using the sea born library, has been generated indicating the relationship of every feature with the price range.

Plots indicating the relationship of ram and battery power with price range have also been generated.

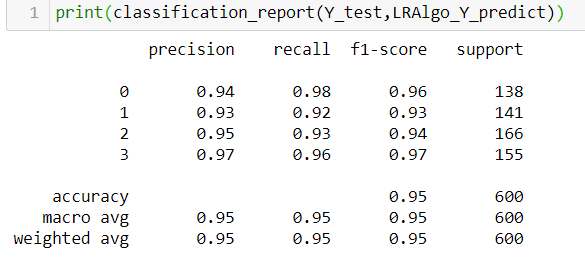


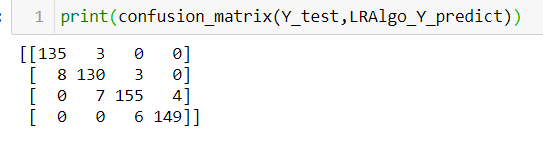


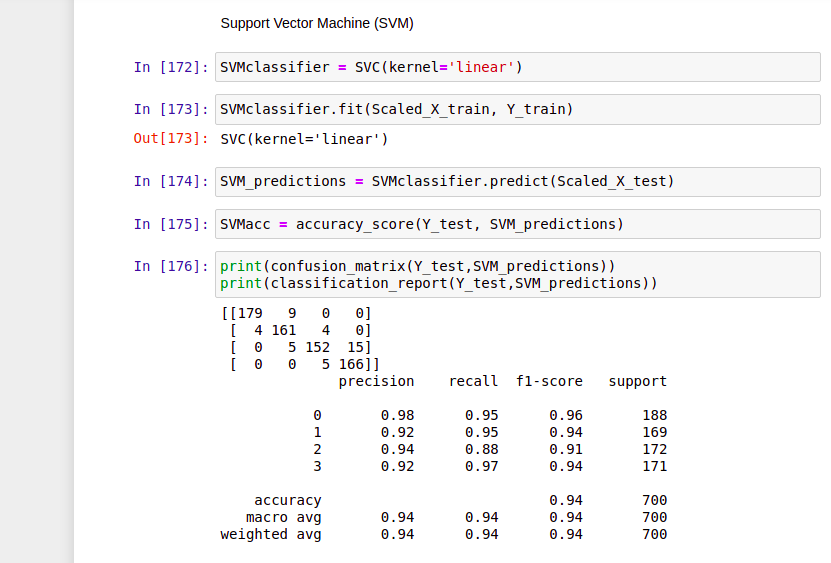
**Classification:**

After performing data visualization, the process of model building started. The training data set consisted of 70 percent of the original data set and the testing set consisted of 30 percent of the original data. Since decision trees are implemented well in the context of variables which are not linearly separable machine decision trees classifier was used. It generated an accuracy of 83 percent. Below is the classification report indicating different evaluation measures such as precision, recall, f1 score, accuracy, etc along with the confusion matrix generated from the classifier.

The second algorithm and third algorithm that are used for this task are Logistic Regression and Support Vector Machine. Logistic Regression is a classification algorithm that generates probabilities using the Sigmoid function and selects the class outcome on the basis of those probabilities. It generated an accuracy of 95 percent. Support Vector Machine is another classification algorithm which is used , it generates an accuracy of 94 percent. The internal working of this algorithm is such that it generates a hyper plane which creates a distinction between the four feature classes. Using this distinction it outputs a prediction. Below are the classification report and confusion matrix generated from these two classifiers.







The fourth and fifth classification algorithms which are being used are XG boost and K Nearest Neighbors. XG boost is an ensemble learning method. Ensemble learning is a methodology in Machine Learning where several predictions of several models are combined to derive an aggregated prediction. K Nearest Neighbors is another classification algorithm which uses the concept of Euclidean Distance to predict outcomes. The user selects the number of neighbors to be used in the prediction and calculates the testing data’s distance with the neighbors. The neighbor with which the testing data has the shortest distance , the algorithm predicts that the data belongs to that particular neighbor’s class. XG boost gives an accuracy of 90 percent and KNN gives an accuracy of 69 percent.

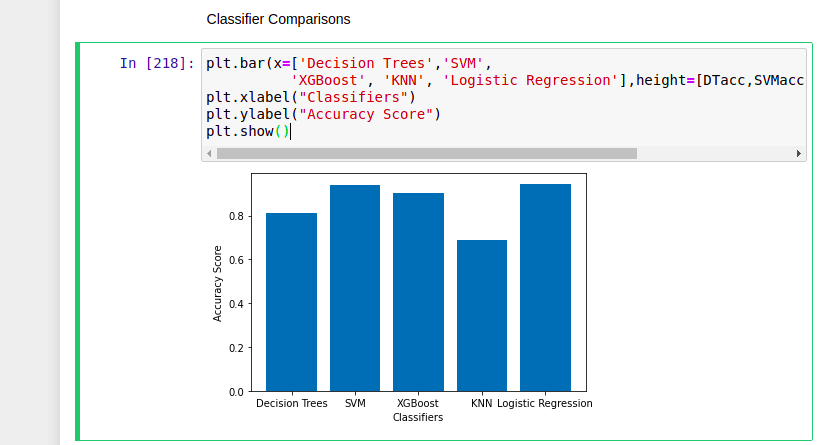
In order to increase accuracy an improvement technique known as standard scaler has been used on the training data and the testing data. Standard scaler standardizes the data by skimming the range of the values of each attribute. As a result, it lessens the standard deviation of the values of all the attributes and lowers the chances of there being any outliers in the data that the algorithm would incorrectly predict.

**Conclusion:**

In conclusion to our task two algorithms have given the best accuracy , Logistic Regression and

Support Vector Machine. Logistic Regression producing an accuracy of 95 percent and Support

Vector Machine producing an accuracy of 94 percent.



|  |  |  |
| --- | --- | --- |
| **Number of Features** | **Classifier** | **Accuracy** |
| 21 | Logistic Regression | 95% |
| 21 | Support Vector Machine | 94% |
| 21 | XG Boost | 90% |
| 21 | Decision Trees | 83% |
| 21 | K Nearest Neighbors | 69% |

**Future Work:**

For the purpose of Mobile price classification many newly developed artificial intelligence

Techniques can also be used such as Deep Learning Neural Networks which have been known

to perform exceptionally on large data sets with high dimensions. Different accuracy

improvement techniques other than standard scaling such as Principal Component Analysis

can also be tested for such a task.

**References:**

1. Kanwal Noor and Sadaqat Jan, “Vehicle Price Prediction System using Machine Learning Techniques”, International Journal of Computer Applications, June 2017.
2. Sameerchand Pudaruth, “Predicting the Price of Used Cars using Machine Learning Techniques”, International Journal of Information & Computation Technology, 2014
3. Limsombunchai, “House Price Prediction: Hedonic Price Model vs. Artificial Neural Network”, New Zealand Agricultural & Resource Economics Society Conference, New Zealand, 2004.
4. Shonda Kuiper, “Introduction to Multiple Regression: How Much Is Your Car Worth?”, Journal of Statistics Education, November 2008.
5. Kaggle.com