Abstract :

In this Modern Era, Smartphones are an integral part of the lives of human beings. When a smartphone is purchased ,many factors like the Display, Processor, Memory, Camera, Thickness, Battery, Connectivity and others are taken into account . One factor that people do not consider is whether the product is worth the cost . As there are no resources to cross validate the price , people fail in taking the correct decision. This paper looks to solve the problem by taking the historical data pertaining to the key features of smartphones along with its cost and develop a model that will predict the approximate price of the new smartphone with a reasonable accuracy. The dataset[12] used for this purpose has taken into consideration 21 different parameters for predicting the price of the phone . Random Forest Classifier, Support Vector Machine and Logistic Regression have been used primarily. Based on the accuracy , the appropriate algorithm has been used to predict the prices of the smartphone. This not only helps the customers decide the right phone to purchase , it also helps the owners decide what should be the appropriate pricing of the phone for the features that they offer. This idea of predicting the price will help the people make informed choice when they are purchasing a phone in the future. Among the three classifiers chosen , Support Vector Machine had highest accuracy of 99%. Further XG boost (Hyper-parameter Tuned) was used to predict the prices of the phone.

Keywords:

Introduction:

Price is the most effective attribute of marketing and business. The very first question of costumer is about the price of items. All the costumers are first worried and thinks “If he would be able to purchase something with given specifications or not”. So to estimate price at home is the basic purpose of the work. This paper is only the first step toward the above mentioned destination.

Artificial Intelligence-which makes machine capable to answer the questions intelligently- now a days is very vast engineering field. Machine learning provides us best techniques for artificial intelligence like classification, regression, supervised learning and unsupervised learning and many more. Different tools are available for machine learning tasks like MATLAB, Python, cygwin, WEKA etc. We can use any of classifiers like Decision tree , Naïve Bayes and many more. Different type of feature selection algorithms are available to select only best features and minimize dataset. This will reduce computational complexity of the

problem. As this is optimization problem so many optimization techniques are also used to reduce dimensionality of the dataset.

Mobile now a days is one of the most selling and purchasing device. Every day new mobiles with new version and more features are launched. Hundreds and thousands of mobile are sold and purchased on daily basis. So here the mobile price_class prediction is a case study for the given type of problem i.e finding optimal product. The same work can be done to estimate real price of all products like cars, bikes , generators, motors, food items, medicine etc.

Many features are very important to be considered to estimate price of mobile. For example Processor of the mobile. Battery timing is also very important in todays busy schedule of human being. Size and thickness of the mobile are also important decision factors. Internal memory, Camera pixels, and video quality must be under consideration. Internet browsing is also one of the most important constraints in this technological era of 21st century. And so is the list of many features based upon those, mobile price is decided. So we will use many of above mentioned features to classify whether the mobile would be very_economical, economical, expensive or very_ expensive.

The structure of the paper is as follows. Next section is problem statements .3 rd Section contains Methodology and Experimental procedure. Section 4 is the summary of the results. Comparative study is done in section 5. After that paper is concluded in section 6. Outcomes of the work are discussed in section 7. At last in 8th section some suggestions about future work are given.

Problem statement :

In the competitive mobile phone market companies want to understand sales data of mobile phones and factors which drive the prices. The objective is to find out some relation between features of a mobile phone(eg:- RAM, Internal Memory, etc) and its selling price. In this problem, we do not have to predict the actual price but a price range indicating how high the price is.

Data Summary :

- We have a data_mobile_price_range data for our analysis and model building.
- This data set contain total rows 2000 & total columns 21.
- This data set contain battery_power, blue, clock_speed, dual_sim, fc, four_g, int_memory, m_dep, mobile_wt, n_cores, pc, px_height, px_width, ram, sc_h, sc_w, talk_time, three_g, touch_screen, wifi, price_range, dtype='object’.

Data Collection :

- Battery_power - Total energy a battery can store in one time measured in mAh.
- Blue - Has bluetooth or not.
- Clock_speed - speed at which microprocessor executes instructions.
- Dual_sim - Has dual SIM support or not.
- Fc - Front Camera mega pixels.
- Four_g - Has 4G or not.

- Int_memory - Internal Memory in Gigabytes.
- M_dep - Mobile Depth in cm.
- Mobile_wt - Weight of mobile phone.
- N_cores - Number of cores of processor.
- Pc - Primary Camera mega pixels.
- Px_height and Px_width - Pixel Resolution Height and width.
- Ram - Random Access Memory in Mega Bytes.
- Sc_h and Sc_w - Screen Height and width of mobile in cm.
- Talk_time - longest time that a single battery charge will last when you are.
- Three_g - Has 3G or not.
- Touch_screen - Has touch screen or not.
- Wifi - Has wifi or not.
- Price_range - This is the target variable with value of 0(low cost),1(medium cost),2(high cost) and3(very high cost).

Dimensionality Reduction:

Dimensionality reduction is the process of reducing the number of random variables(Features) under consideration, by obtaining a set of principal variables. The higher the number of features, the harder it gets to visualize the training set and then work on it. Sometimes, most of these features are correlated, and hence redundant. This is where dimensionality reduction algorithms come into play. Two types of Dimensionality reduction algorithms are there ie Feature selection, Feature extraction.

Feature Selection :

In feature selection we are interested in finding k of the d dimensions that give us the most information, and we discard the other (d − k) dimensions.

Feature Extraction:

In feature extraction we are interested in finding a new set of k dimensions that are combinations of the original d dimensions for example Principal Component Analysis. Here feature selection algorithms are used. There are two approaches: Forward selection and backward selection.

Forward Selection:

In forward selection, we start with no variables and add them one by one, at each step adding the one that decreases the error the most, until any further addition does not decrease the error (or decreases it only slightly).

Backward Selection:

In backward selection we start with all variables and remove them one by one, at each step removing the one that decreases the error the most (or increases it only slightly), until any further removal increases the error significantly.

Exploratory data analysis:

In statistics, exploratory data analysis (***EDA***) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily ***EDA*** is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

Correlation of independent variable with target variable :

- RAM has strong positive correlation with the Price_range. and we know that Mobiles with high RAM are very costly. Thus

RAM increases price range also increase.

- Battery_power also has positive correlation with the price range.Generally mobiles having high prices comes with good battery power.
- primary camera mega pixels and front Camera mega pixels have correlation (it make sense because both of them reflect technology level of resolution of the related phone model) but they do not effect price range.
- having 3G and 4G is somewhat correlated, Nowdays most of the smart mobiles has both type of options. This could be the reason that they are correlated.
- sc_h and sc_w are positively correlated.
- there is no highly correlated inputs in our dataset, so there is no multicollinearity problem.

Algorithm :

Decision Tree :

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

Random Forest :

A random forest is a supervised machine learning method built from decision tree techniques. This algorithm is used to anticipate behaviour and results in a variety of sectors, including banking and e-commerce.A random forest is a machine learning approach for solving regression and classification issues. It makes use of ensemble learning, which is a technique that combines multiple classifiers to solve complicated problems.A random forest method is made up of a large number of decision trees. The random forest algorithm's 'forest' is trained via bagging or bootstrap aggregation. Bagging is a meta-algorithm ensemble that increases the accuracy of machine learning algorithms.The outcome is determined by the (random forest) algorithm based on the predictions of the decision trees. It forecasts by averaging or averaging the output of several trees. The precision of the outcome improves as the number of trees grows.

KNN Classifier:

The K Nearest Neighbor method is a type of supervised learning technique that is used for classification and regression. It's a flexible approach that may also be used to fill in missing values and resample datasets. K Nearest Neighbor examines K Nearest Neighbors (Data points) to forecast the class or continuous value for a new Datapoint, as the name indicates.

SVM Classifier:

Support Vector Machine, or SVM, is a prominent Supervised Learning technique that is used for both classification and regression issues. However, it is mostly utilised in Machine Learning for Classification purposes.The SVM algorithm's purpose is to find the optimum line or decision boundary for categorising n-dimensional space so that we may simply place fresh data points in the proper category in the future. A hyperplane is the optimal choice boundary.

XG Boost:

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

Logistic regression:

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.

Conclusion:

In this article, we looked at classification. Classifiers represent the intersection of advanced machine theory and practical application. These algorithms are more than just a sorting mechanism for organising unlabeled data instances into distinct groupings. Classifiers include a unique set of dynamic rules that include an interpretation mechanism for dealing with ambiguous or unknown values, all of which are suited to the kind of inputs being analysed. Most classifiers also utilise probability estimates, which enable end-users to adjust data categorization using utility functions.

This work can be concluded with the comparable results of both Feature selection algorithms and classifier. This combination has achieved maximum accuracy and selected minimum but most appropriate features. It is important to note that in Forward selection by adding irrelevant or redundant features to the data set decreases the efficiency of both classifiers. While in backward selection if we remove any important feature from the data set, its efficiency decreases. The main reason of low accuracy rate is low number of instances in the data set. One more thing should also be considered while working that converting a regression problem into classification problem introduces more error.

Outcomes Of The Work :

- ✔ Cost prediction is the very important factor of marketing and business. To predict the cost same procedure can be performed for all types of products for example Cars, Foods, Medicine, Laptops etc.
- ✔ Best marketing strategy is to find optimal product (with minimum cost and maximum specifications). So products can be compared in terms of their specifications, cost, manufacturing company etc.
- ✔ By specifying economic range a good product can be suggested to a costumer.

References:

Ethem Alpaydın, 2004. Introduction to Machine Learning, Third Edition. The MIT Press Cambridge, Massachusetts London, England

Mariana Listiani , 2009. “Support Vector RegressionAnalysis for Price Prediction in a Car LeasingApplication”. Master Thesis. Hamburg University of Technology.

Introduction to dimensionality reduction, A computer science portal for Geeks.https://www.geeksforgeeks.org/dimensionalityreduction/ (Last Accessed on Monday , Jan 201822, 3 PM)

Thu Zar Phyu, Nyein Nyein Oo. Performance Comparison of Feature Selection Methods. MATEC Web of Conferences42, (2016).