Udacity Machine Learning NanoDegree

Mobile Price Classification

Madhu Babu

March 2nd ,2019

1.Definition

Project Overview:

Our dataset is mobile dataset.The mobile phones boackground is discussed in the following.

In the present society,business field is more demanding than any other field.Here we take the data regarding this field.That is here we predict mobile prices according to the features of the mobile.A mobile phone is a portable telephone that can make and receive calls over a radio frequency link while the user is moving within a telephone service area. The radio frequency link establishes a connection to the switching systems of a mobile phone operator, which provides access to the public switched telephone network . Modern mobile telephone services use a cellular network architecture, and, therefore, mobile telephones are called mobile phones or cell phones. In the present world, mobile phones support a variety of other services, such as text messaging, MMS, email, Internet access, short-range wireless communications (infrared, Bluetooth), business applications, video games, and digital photography. Mobile phones offering only those capabilities are known as feature phones; mobile phones which offer greatly advanced computing capabilities are referred to as smartphones.These can be offer many features which increses the range of the mobile prices.

Out dataset was derived to predict the prices range for the mobiles.This dataset is taken from kaggle. The mobile dataset contains 2000 datapoints with 22 attributes.The mobile database contains the attributes are battery_power,blue,clock_speed,dual_sim etc.Our main goal is to predict the given mobile price.Here we need to estimate the price_range that is how high the mobile price is.We don't need exact price of the mobile.So,here we are supposed to train a classification model that predicts the price ranges by using the features of the mobiles.

The prices range of the mobiles should be of four classes i.e., 0,1,2,3.The 0 price range indicates that the price is low ,1 indicates moderate price range 2 indicates that the price range is high and 3 indicates the price range is very high.

I have seen many websites,I find one interesting academic research of Muhammad Asim and Zafar Khan that has been published on this mobile price classification.The following is the reference link for the research

https://www.researchgate.net/profile/Muhammad_Asim41/publication/323994340_Mobile_Price_Class_prediction_using_Machine_Learning_Techniques/links/5b2b23b94585150c63446830/Mobile-Price-Class-prediction-using-Machine-Learning-Techniques.pdf

Problem Statement: This project is about the Mobile Database. A man started his own mobile company.He doesnot know how to estimate the mobile prices.So he collected sales data of various mobile companies .He wants to find some relation between features of mobile and the selling price of the mobile.So he choose you as his employee.Our work is to predict the rance of mobile price but not the exact price of the mobile.

Hence this project comes under classification.So we should apply a classification model to the dataset so that we can get the best model to predict the range of mobile price . Furthur more the model should be very best useful for the real world applications. By training model with the data,we can predict for what features the predictions are varied and by the different

datapoints,we can also predict that which features are mainly effecting the target variable price_range.We can also find the correlation between those features among themselves.

The size of the data sources are impossible for the human analyst to come up with the interesting information that will help you in the decision making process. So, Data mining models will completely help in the performance of these campaigns.

The purpose is to judge effectively by identifying the main characters that effect the success based on handful of algorithms that will test the following algorithms like Logistic Regression,KNN, Decision Trees, Random Forest etc..we the experimental results will tell the performance of the models statistical metrics and then we can be able to say which model correctly judges the decisions in predicting the mobile prices.

The dataset is taken from the kaggle.

Metrics: The evaluation metric used in our learning algorithm and the benchmark model is accuracy score. The accuracy score decides which model is best required for the prediction of the mobile price ranges.The fbeta score can also can be used but we are not taking fbeta score because there are four different classes for the target variable for our project and the number of each classes in the whole dataset are equal i.e, the classes 0,1,2 and 3 having same number of instances in the dataset.The accuracy_score can be imported from sklearn.metrics.

II. Analysis

Data Exploration: This project is about the Mobile Database. The mobile dataset contains 2000 datapoints with 22 attributes.The mobile database contains the attributes are 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,s c_w,talk_time,three_g,touch_screen,wifi,price_range.Here we wants to find some relation between features of mobile and the selling price of the mobile. Our work is to predict the range of mobile price but not the exact price of the mobile. Hence this project comes under classification .So,to solve this problem, this model came into

existence.This model tells the particular mobile price range based on the features of that phone.

In order to judge the mobile price range this dataset ,dataset makes some of the features which best justifies the properties of mobiles.The dataset contains the following attributes.

features:

id:ID

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:Pixel Resolution Height

px_width:Pixel Resolution Width

ram:Random Access Memory in Megabytes

sc_h:Screen Height of mobile in cm

sc_w:Screen 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

Target variable:

price_range:mobile price range i.e 0(low price),1(medium price),2(high price),3(very high price)

Here the target variable is price_range which we have to predict.Remaining all are features.

From the above features,we can say that the high feature quality will results the high quality mobile which is having a high cost.We can train the model using the data provides features and target.

Because of known underlying concept structure, this database may be particularly useful for testing constructive induction and structure discovery methods.Hence ,so we can accurately predict the correct judgements as the features are very well known and also by these features,we can make approximately best decisions which makes judgement that our model is working well or not.

statistical description:

>Total number of training records : 2000

>The total number of records belong to class 0:500

>The total number of records belong to class 1:500

>The total number of records belong to class 2:500

>The total number of records belong to class 3:500

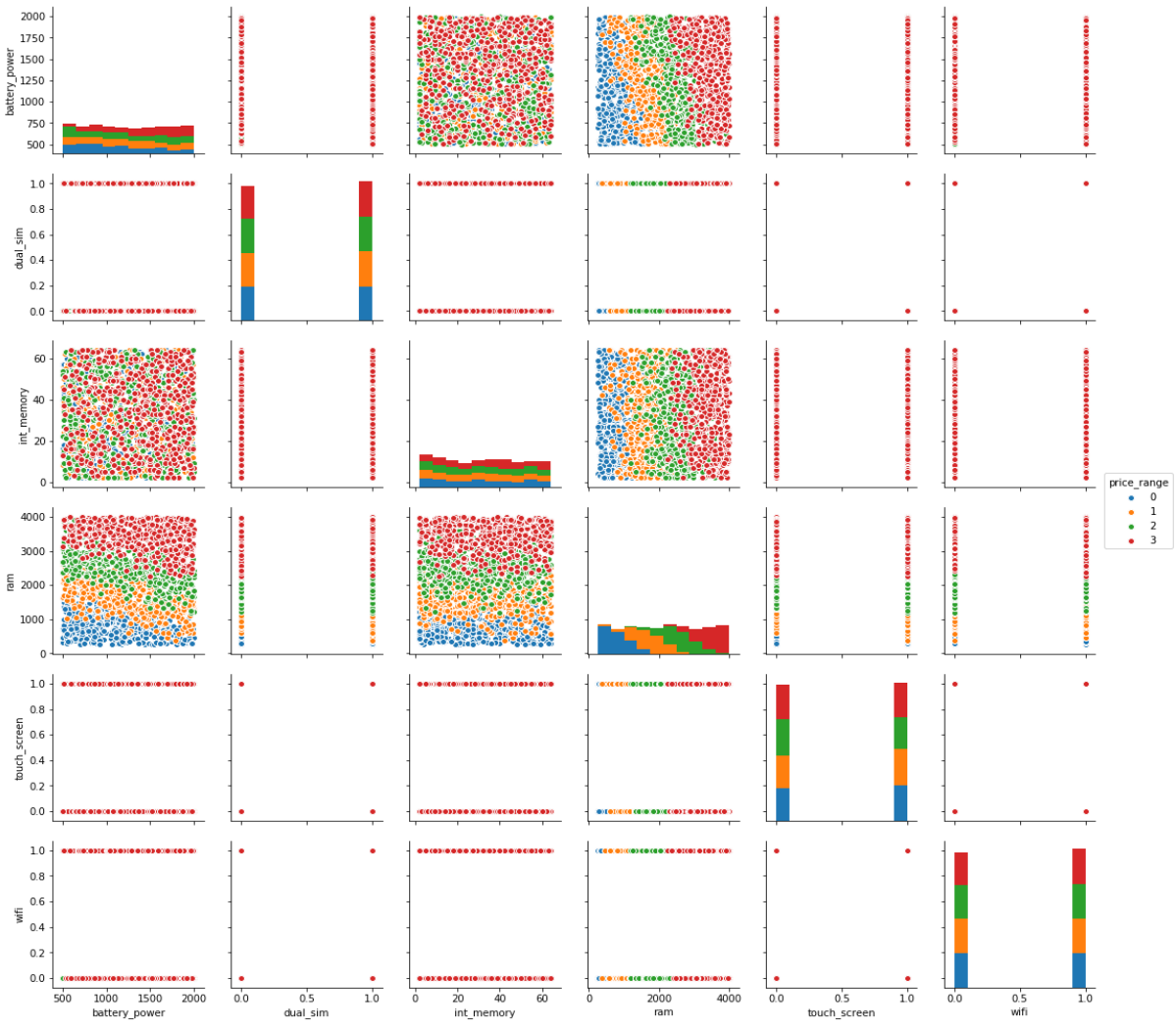
From the above records ,we can say that the data is balanced.The following table describes the statistical description of the mobile dataset attributes.

	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	...	
count	2000.000000	2000.0000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	...	2
mean	1238.518500	0.4950	1.522250	0.509500	4.309500	0.521500	32.046500	0.501750	140.249000	4.520500	...	6
std	439.418206	0.5001	0.816004	0.500035	4.341444	0.499662	18.145715	0.288416	35.399655	2.287837	...	4
min	501.000000	0.0000	0.500000	0.000000	0.000000	0.000000	2.000000	0.100000	80.000000	1.000000	...	0
25%	851.750000	0.0000	0.700000	0.000000	1.000000	0.000000	16.000000	0.200000	109.000000	3.000000	...	2
50%	1226.000000	0.0000	1.500000	1.000000	3.000000	1.000000	32.000000	0.500000	141.000000	4.000000	...	5
75%	1615.250000	1.0000	2.200000	1.000000	7.000000	1.000000	48.000000	0.800000	170.000000	7.000000	...	9
max	1998.000000	1.0000	3.000000	1.000000	19.000000	1.000000	64.000000	1.000000	200.000000	8.000000	...	1

8 rows × 21 columns

Exploratory Visualization:

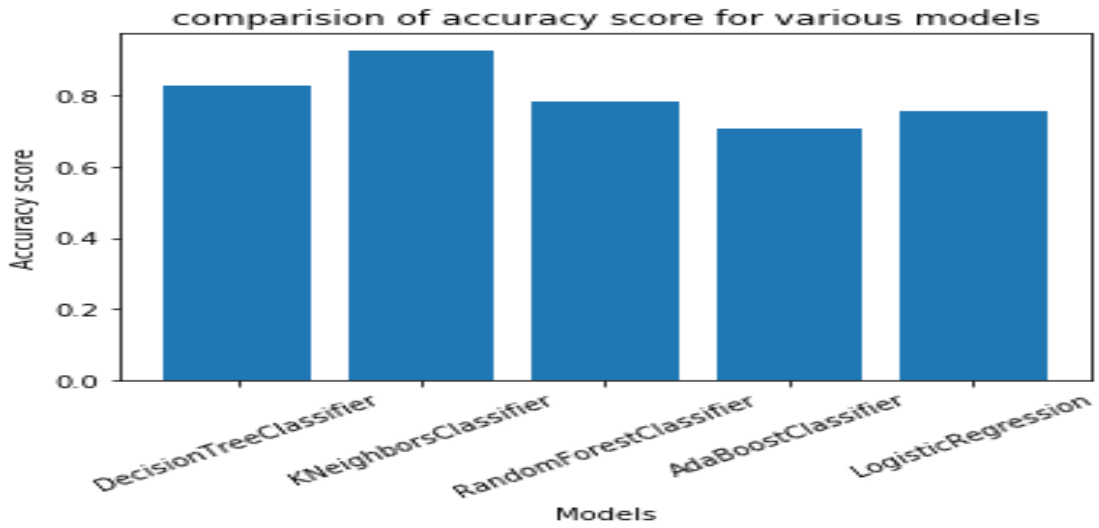
Graph1:



The above pairplot graph represents the datapoints of different features.We have taken 6 features.So total 36 graphs can be shown in the figure.From this pairplot graph,we have

found that there is no corelation between the features and also there are no outliers found.So we don’t need to apply any scaling methods.

Graph2:



The above bar graph shows the accuracy scores of different classifiation models. The appliedmodels are given below.

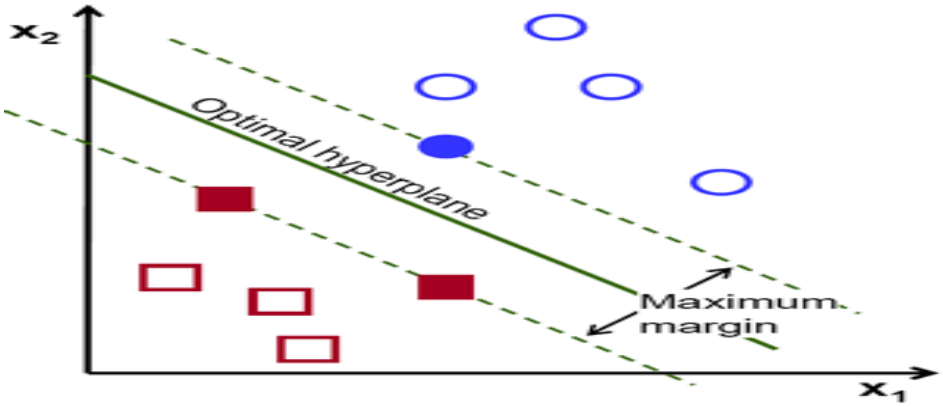
DecisionTreeClassifier,KneighborsClassifier,RandomForestClassifier,AdaBoostClassifier,LogisticRegression. Here we can easily findout the best model which gives best accuracy score.

Algorithms and Techniques:

For the above mobile database,the following algorithms can be used to get the best model that it can give the best predictions or decisions in which the it best describes the range of mobile price for the given test data.They are KNN(KNeighbours Classifier),SVM,Random Forest Classifier,Logistic Classifier,Decision Tree Classifier,Adaboost Classifier.

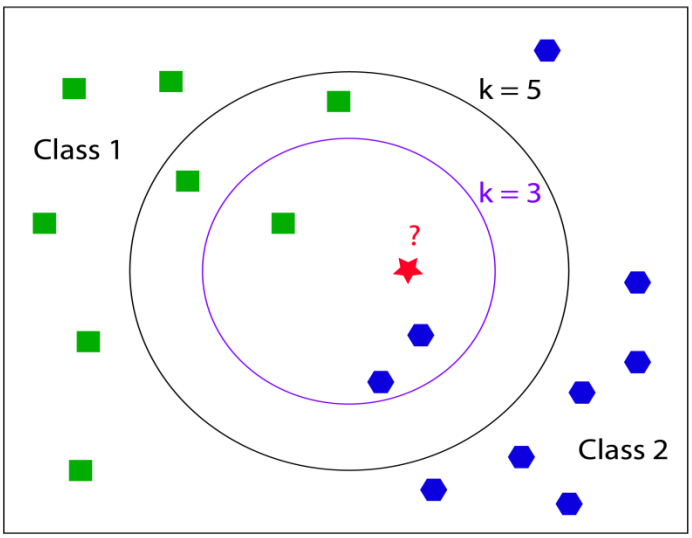
Support Vector Classifier:

Support vector machine (SVM) is a supervised machine learning algorithm which can be used for classification.SVM is used for image based gender identification. SVM is effective when there is a lot of features or in high dimensional space.SVM makes use of hyperplane which acts like decision boundary between various classes.It works well very well with a clear margin of seperation between support vectors.



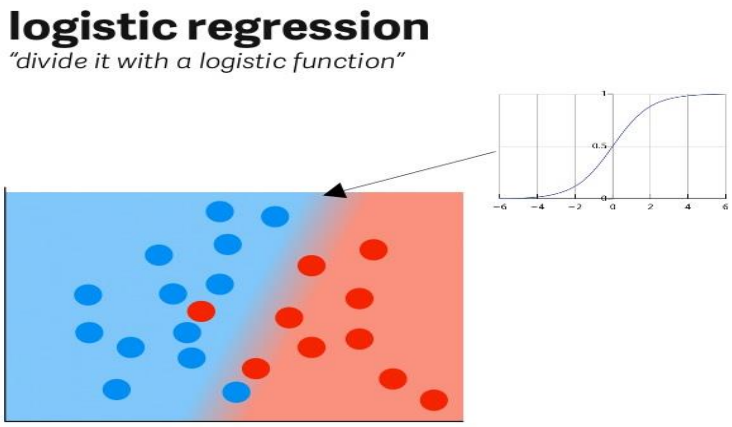
k-Nearest Neighbors:

K-Nearest Neighbors is one of the basic classification algorithms in Machine Learning. It can be used in pattern recognition and intrusion detection. It is widely usable in real-life scenarios since it is non-parametric. It does not make any underlying assumptions about the distribution of data.

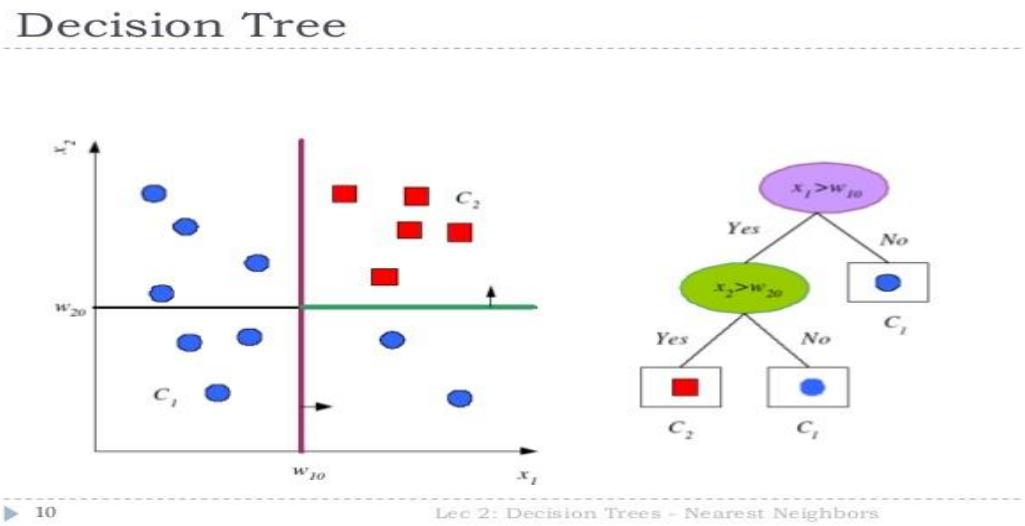


Pseudo code:Load the training and testing data. Choose a k value. For each point in the sample, Calculate the Euclidean distance to all the training points. Choose k points that are in less distance from our sample point. Assign a class to our sample point based on the majority of classes for the chosen points

Logistic Regression: Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. It is used for binary classification. It predicts the probability of occurrence of an event by fitting data to a logistic or sigmoid function.



Decision Trees: A decision tree is a predictive model based on a branching series of Boolean tests that use specific facts to make more generalized conclusions. Moreover, It is a graphical representation of specific decision situations where each node represents a feature (attribute), each link (branch) represents a decision (rule) and each leaf represents an outcome (categorical or continues value). The whole idea is to create a tree for the entire data and process a single outcome at every leaf.



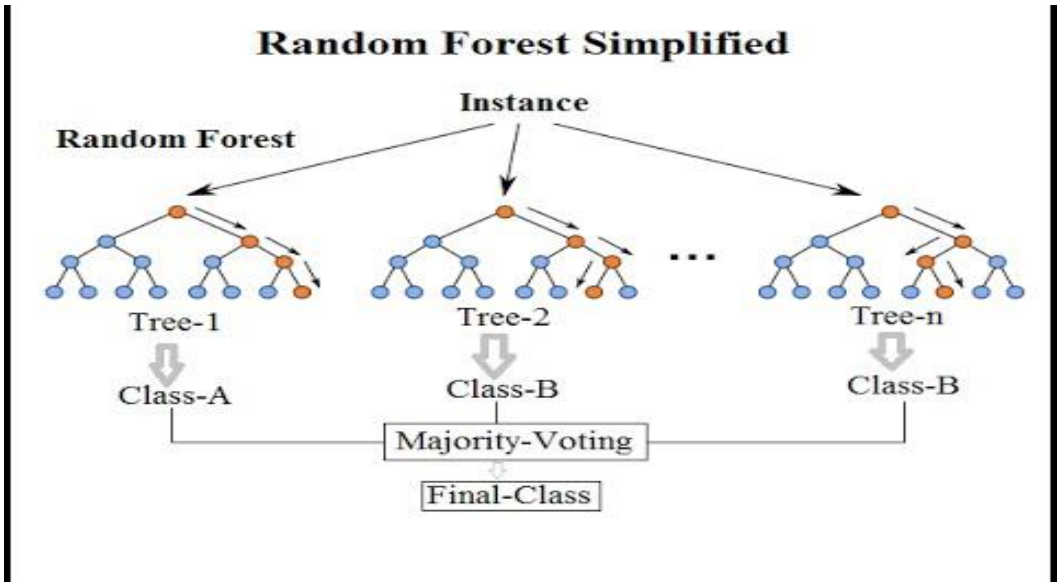
Random Forest: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. In Random Forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. You can even make

trees more random, by additionally using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree does).

In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results .

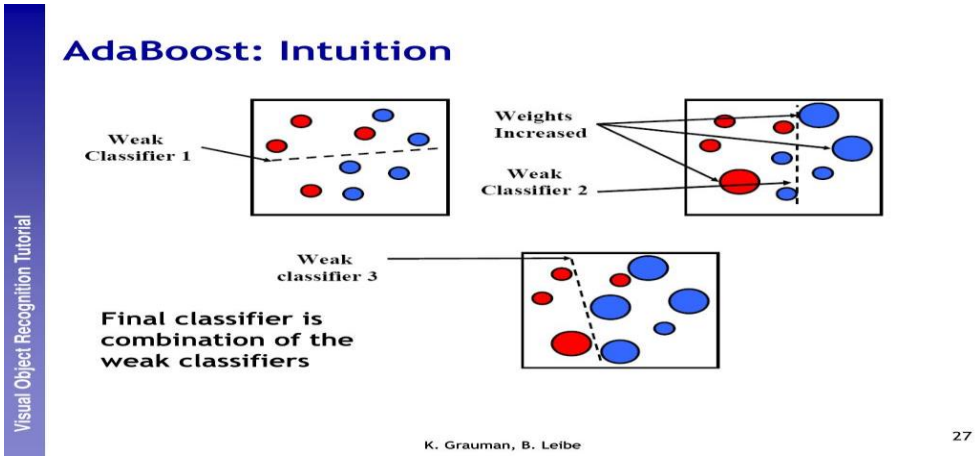
Pseudo code: Randomly select k features from total m features. Among the k features, calculate the root node using the best split point. Split the node into daughter nodes using the best split. Repeat 1 to 3 steps until 1 number of nodes has been reached. Build forest by repeating steps 1 to 4 for n number times to create n number of trees

For prediction: Takes the sample data and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target). Calculate the votes for each predicted target. Consider the high voted predicted target as the final prediction from the random forest algorithm .



AdaBoost classifer: AdaBoost classfier is an boosting algorithm.It is an ensemble method that creates a strong classifier from a number of weak classifiers. This is done by building a model from the training data, then creating a second model that attempts to correct the errors from the first model. Models are added until the training set is predicted perfectly or a maximum number of models are added.

It fits a sequence of weak learners on different weighted training data. It starts by predicting original data set and gives equal weight to each observation. If prediction is incorrect using the first learner, then it gives higher weight to observation which has been predicted incorrectly. Being an iterative process, it continues to add learners until a limit is reached in the number of models or accuracy.



Mostly, we use decision stamps with AdaBoost. But, we can use any machine learning algorithms as base learner if it accepts weight on training data set. We can use AdaBoost algorithms for both classification and regression problem.

Benchmark:

As we said in the capstone proposal, we taken the bench mark model as SVC which was derived from sklearn.svm. Hence the dataset takes this model as benchmark model,the results of this model would be related to the results of the final model which tells us the exact class of the mobile price.It is because our benchmark model results should be far more better than the results of the learned algorithm,then we can perfectly decide that our project is predictable or not.

The results for this model is

The accuracy score of this model is 0.23

Hence the value of the final learned algorithm should be far better than this value.

III. Methodology:

Data PreProcessing:

In our Mobile Dataset,all the attributes are numerical values.In the step of preprocessing, we don’t need to do any scaling method because from the above pairplot we have seen that there is no outliers for our data.And also we don’t have to change any values because we have all attributes are numerical values.

Implementation:

The following is done in the implementation process:

>First, we explored the dataset befor applying any machine learning techniques then we came to know some information like how the data is distributed and whether we need to preprocess the data or not as described in above Data exploration and Data visualization sections.

>Next, we split the preprocessed dataset into two sets as training set (80%) and testing set (20%).

> For each algorithm we considered, we fit the training set to the model and tested the model against testing set.

>Finally, we compared the model by using accuracy_score.

>The model with the highest accuracy_score is selected.

While performing I find some difficulty for choosing the number of k_nearest neighbors.Later I put a list of number of k_nearest neighbors for performing GridSearchCV.The following is the code used in this scenario.

```
param_grid = {'n_neighbors' : [5,7,10] }

grid_obj=GridSearchCV(dtc,param_grid,cv=5,scoring='accuracy')
```

The accuracy scores of all models are listed below to find the best model to perform GridSearchCV.

>The scores as follows:

The Betchmark model(SVM) has an accuracy score as 0.23

For DecisionTreeClassifier,the accuracy score is 0.83

For KNeighborsClassifier,the accuracy score is 0.9275

For RandomForestClassifier,the accuracy score is 0.785

For AdaBoostClassifier,the accuracy score is 0.705

For LogisticRegression,the accuracy score is 0.7575

Here KneighborsClassifier accuracy score is high, we choose this model to solve our problem.

So we apply GridSearchCV on KneighborsClassifier by giving the values n_estimators=[5,7,10] ,from this we can get the best results by trying the different values for number od nearest neighbors.

scoring=’accuracy’ because we choose the metric as accuracy score.

cv=5 the cross validation generator iterates our model by 5 times to get the best optimized model which gives more accuracy score than the bench mark accuracy score.

IV. Results:

Model Evaluation and Validation:

The final model has the parameters as follows:

>n_neighbors=[5,7,10]

> cv=5

>scoring='accuracy'

> The final model i.e., optimized model performed well, Since there is a considerable increase in the accuracy score.

Robustness:

We can train our model with all the features available and test them against the testing data.Then we get the results as follows:

	acacuracy score on tesing data
unoptimized mode	0.9275
optimized model	0.9575

That is the final model performs well as it predicts the testing data more accurately. So we can say that the final model is robust and we can trust the results from this model.

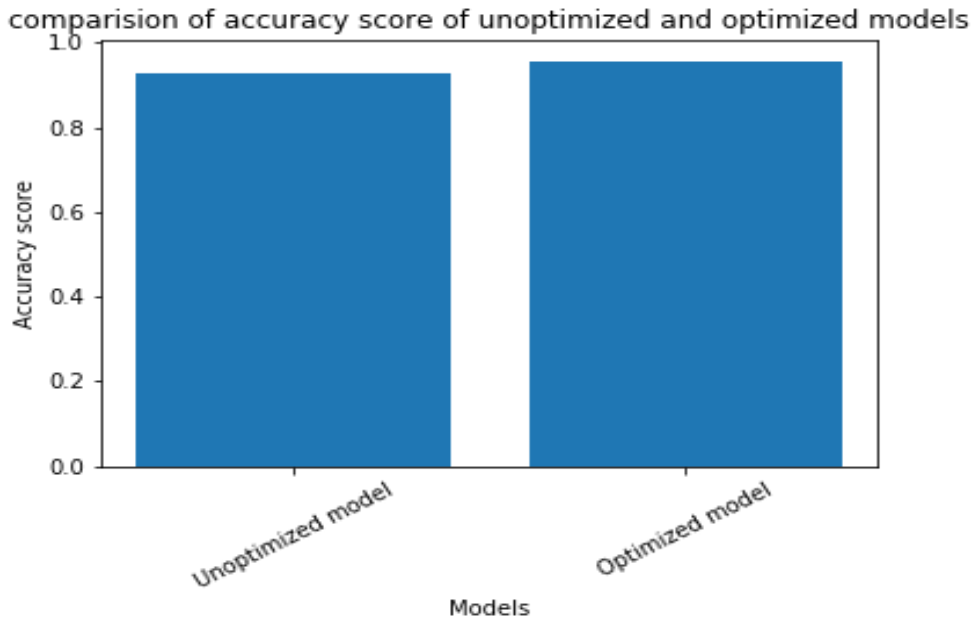
Justification :

Final accuracy score on the testing data: 0.9525

The final model's scores are good when compared to Bench Mark model's scores. So, we can consider our final model as new Bench mark model if any new classification algorithm wants to be applied.So from here, the model can best predict mobile price range i.e.,0,1,2,3.This solves our problem.

V.Conclusion:

Free-form visualization:



From this bar graph,we can easily find that the optimized model gives more accurate results than unoptimized model because optimized model has high accuracy score than that of unoptimized model.

Reflection:

The entire end-to-end problem solution is as follows:

First we understand the given problem by Domain Background and realize how the problem raised.

Exploring the data:

We imported necessary packages required for solving the problem. We loaded the dataset. Performed statistical operations.

Data Preprocessing:

Dividing the data into training set and testing set.

>Model Selection:

Applied various algorithms on the dataset.

- SVM
- Decision trees

- Adaboost
- K-Nearest Neighbors
- Random Forest
- Logistic Regression

> Compared each algorithm with the Benchmark Model using the metrics accuracy_score .

> And we select the best model among them.

Model Tuning:

> Tuned the selected model with different combination of parameters so that the model performs well.

Interesting aspect:

The interesting aspects of this project is visualization.Through visualization we can easily analyze the data or outputs more accurately than through looking for values.

I found no difficulty to do this project because our dataset is balanced and there is no correlation between features of this model and we need not do procedures like scaling,.

Improvement:

- In gridsearch operation, we can add more parameters to improve the performance.
- I think this problem can have another best solution if the final model is tuned deeply.
- We can provide more classification algorithms to do this task.Because we used some of these classification algorithms for our task.There may be a chance that other classification algorithm works well than our classification model.