Technical Documentation for the Classification ML Project

By *Saraswat Mukherjee*

Introduction:-

There are several mobile phones nowadays available in the market, available in different price range. The price range depends usually depends on many factors like model, features etc. The mobile with more features and more attractive looks has been seen to fall in greater price as compared to those with lesser features. So we need to analyse all the features and on its basis we can predict the price range of the mobiles.

Abstract:-

We have got a dataset with several features like battery power, clock speed, internal memory etc. We will be observing and analysing the features along some other stuffs like wrangling, EDA and finally implementing machine learning algorithms onto it. By creating a proper machine learning models along with reconciling their accuracies we will be predicting the price range of the mobiles.

EDA:-

Data Collection:-

Now we will start our journey of working out with the dataset. Firstly we will import all the libraries which will be required in our journey. Along the journey, while executing our codes we may get few warnings. So we will also import and use a warning library to take care of the warnings.

Then we will import the dataset and observe it. We will try to get the non null count for each column which represents the features of the mobiles and their respective data types. We will try to now get the count, mean standard deviation, minimum, maximum and percentiles of it. This will provide an extensive and a wide idea about the distribution of the dataset. While analysing the data we found a couple of anomalies. We found a couple of rows with 0 pixel height and 180 rows with 0 screen width which is of course impossible. These missing values consists of 9% of the entire dataset which is quite significant hence can’t be dropped. So we need to replace it by null values which will make it less scattered which will be further be replaced by mean and nearest possible values. Now we can say our dataset is free of anomalies or null values.

Outlier Detection:-

At times we have outliers in our dataset which makes the data skew and scattered. It may create few problems later while working with the data. So getting rid of the outliers is very important in order to work peacefully with our data. So firstly we need to detect the outliers for each feature of the data by representing them as a box plot. Fortunately in this dataset we did not find any extreme outliers.

Data Analysis:-

There are many features of mobiles we have as per our dataset. Combinations of these features with different patterns define the price range of mobile phones. So we need to check how each feature is varying and connected to the price range of the mobiles. We will put them in the form of point plots to get an idea about the relationship between those features and price range. Then, we will analyse all the univariate and multivariate features separately to have an idea about the availability of few features in the mobiles like presence of wifi, 4G, Bluetooth and touch screen etc. At last we will check the correlation and covariance among all the features with each other. From this we observed that features like 4G and 3G, front camera and rear camera, px height and px width are moderately correlated. But ram and price range are highly correlated with each other which further concludes mobiles with high ram will be comparatively more expensive as compared to those with lesser ram.

Machine Learning:-

Feature Engineering:-

We have features called screen width and screen height which we formulate in certain manner into a diagonal element will be called ‘screen’. Similarly we will formulate two features named px height and px width into a new feature called ‘ppi’. Then we will drop the original features and keep the newly created ones. Then we will check the correlation of the dataset which showed us a negative correlation between ‘screen’ and ‘ppi’.

Feature Importance:-

Here we will slice our data first as per the requirement and assign it to X and Y variable. Then we need to split the X and Y into train and test data separately. We also need the data in proper form and scale to work with it so we will scale and standardise our data by transforming it. Now we will apply our first machine learning algorithm on the data which is Logistic Regression. We will create an instance of Logistic Regression and fit the scaled train data into it. Then we will sort the data frame by descending value of coefficients according to the importance of attributes in it. This way we will try to visualise the feature importance by the values of their coefficients.

Dropping non relevant columns:-

There few features like ‘wifi’, ‘blue’, ‘m_deep’, ‘touch_screen’ which are not much relevant in our data to work with. So it will better to drop them to make our data more crisp and useful. Then we will slice the remaining data and assign them to X and Y variable. It will followed by splitting the X and Y into train and test data each. Then we will scale and standardise the X train and test data. We have now got a filtered data after dropping multiple columns so we will confirm feature importance in the similar way like we did earlier with our newly filtered data this time and check its correspondence.

Model Selection:-

We will work with four machine learning models and apply them on the dataset to create our machine learning classification models and will compare them with each other. Firstly we will go ahead and create functions to visualise the confusion matrix and AUC-ROC curve along with another function to print the accuracy score. These functions will later help us to compare different machine learning models.

Logistic Regression:-

We will now start our journey of model creation by creating our first ML classification model with Logistic Regression. After creating an instance of it we will fit the X and Y train data into it. Then we will form an evaluation matrix which in return will show us the accuracy, confusion matrix and ROC AUC score. Then we will call the earlier created functions to visualise the confusion matrix, AUC-ROC curve and print accuracy score our model’s train and test data. In this classification model we observed that prediction of price range of very high and low cost is excellent, medium and high cost is good. Collectively we can say prediction accuracy is very good but not excellent. Now we will do a hyper parameter tuning and cross validation of the data. There the data will be fit to grid search and try to get the best accuracy score. We will create a cross validation matrix with accuracy, confusion matrix and AUC-ROC score and will call those functions to their visual representation. Now we can see that the overall accuracy score has improved to excellent levels. Prediction accuracy for price range 1 and 2 has increased to excellent levels too. Accuracies can be also further increased by increasing the number of iterations.

Random Forest Classification:-

The next machine learning algorithm we will use is random forest classification. We will create an instance of this algorithm and fit X and Y train data into it to create our model.

Now we will check it with feature importance and visualise it by coefficient values. Then we will try to get an evaluation matrix which will give us accuracy, confusion matrix and ROC-AUC score of for each train and test data. Here we can see that the accuracy is 1 which clearly suggests that the random forest classification model is over fitting. Hence, it is needed to be dropped.

KNN Classification:-

We will create an instance of KNN classifier and fit X and Y train data into it. Then we will get the matrix evaluation matrix of our KNN Classification model like we did with previous models. This will provide us accuracy, confusion matrix and ROC-AUC scores for each train and test data. We will try to get the confusion matrix and ROC-AOC curve in visual form. From this model we observe that prediction accuracy here is less than optimized that of logistic regression and AUC-ROC score is better than logistic classification on average for all 4 price range. Now we will do hyper parameter tuning of this KNN Classification model with grid search. Then we will find the evaluation matrix of the newly tuned KNN model which further contains new accuracy, confusion matrix and AOC-ROC scores for train and test data each and will represent the last two in visual form. Now we can see that after optimisation KNN has improved a lot but multiclass price range for price range=2 is still less than that of Logistic regression model.

SVM Classification:-

The way we did for previous we will first create an instance of svc and fit our X and Y train data into it. We will now create an evaluation matrix of the model to get accuracy, confusion matrix and ROC-AUC scores of the train and test data each. We will also try to visualise the confusion matrix and ROC-AUC curve understand it better. The model here seems over fitting but it can be fixed by optimization. We will do the hyper parameter tuning of this SVM model and try to get the evaluation matrix of the tuned data. As usual, the matrix will show us the accuracy, confusion matrix and ROC-AUC score of train and test data each along with the visualisation of ROC-AUC curve and confusion matrix. This reduced the over fitting and it’s been observed that the prediction of all 4 price range is good.

Conclusion:-

- There we few features which were inter-related so we created new features through feature engineering.
- We observed that ram and battery power has the highest impact on price range of the mobiles.
- Logistic and SVM gave similar accuracy.
- Logistic Regression Classification showed best results after hyper parameter tuning with train accuracy 91.5% and test accuracy of 89.2%.

- SVM also showed the best results after hyper parameter tuning with train accuracy of 91.6% and test accuracy of 89.2%
- Random forest model was over fitting hence of no use.
- After optimisation KNN did really well but for multi class price range prediction of price range = 2 was low as compared to that of Logistic