STA6704-19Spring 0001

Final project report

Mobile price classification

Mithun Mohanraj

[3]

Abstract

The work presented in this report features my final term project upon classifying mobile phones into a price range by performing exploratory data analysis and building various classification models and comparing their performances based on various evaluation metrics.

1.Introduction

Data mining is a computer assisted process for extracting useful information and patterns of enormous data[1]. It is a whole process of collecting, preparing, and analyzing data to extract some useful knowledge which can be applied for decision making for business developments[1]. It acts as a tool to assist humans for making decisions and identify the future trends in business.

In this project, Data mining techniques are applied for finding the price range of mobile phones based on their different features. We propose several models for the task and aims to find the best model based on its performance on new data.

1.1 Project Significance

Mobile phones are everywhere, it is rare to find someone without a mobile phone in this smartphone era. With the advancement in technology, variety of mobile phones are produced and there is a raising need for manufacturers to determine the ideal price for each item. The significance of this project is to explain how data mining techniques are useful for finding the ideal price range for mobile phones given the data.

1.2 Project Application

Mobile price classification can be applied in different contexts like a mobile phone selling company trying to make pricing decisions, Users trying to find a appropriate mobile, web stores like amazon, e-bay, Best-buy to offer deals, finding trends in mobile phone price. The method we propose to classify the mobile phone can be applied to other products as well.

1.3 Project objective

The objective of this project is to propose different data mining techniques for classifying the mobile phones into a price-range and finding the best model based on its ability to classify new mobile phones accurately.

2. Data Exploration

Data exploration helps us understand the data very well. [2]It is discovering and uncovering the underlying patterns in the raw data. we basically find the statistics of data like number of observations, number of observed variables, relationship between observations, relationship between variables and characteristics. We also explore the data through various data visualization tools like scatter plots, histograms, box-plots and pie-charts to define the problem of interest.

2.1 Data source and description

Data is obtained from **kaggle- Mobile price classification.** It contains **2000 observations of 21 Attributes** of various mobile phones. It is actually the sales data of mobile phones of various companies in the market.

2.2 Target problem

The target problem is classifying each mobile phones into a given price range based on the features they include. Target variable Price_range consist of four price ranges- 0(very-low),1(low),2(medium),3(high).

Table 1: Features and their types

Features	Type
battery_power	Num
Ram	Num
Clock_speed	Num
Mobile_wt	Num
Dual_sim	Factor
Int_memory	Num
N_cores	Factor
wifi	Factor
Price_range	Factor

2.3 Descriptive Statistics

[4]Descriptive statistics of data is quantitatively summarizing the individual features present in the data. It is finding statistical measures like mean, median, standard deviation, Maximum value, Minimum value etc.

Table 2: Descriptive statistics of attributes

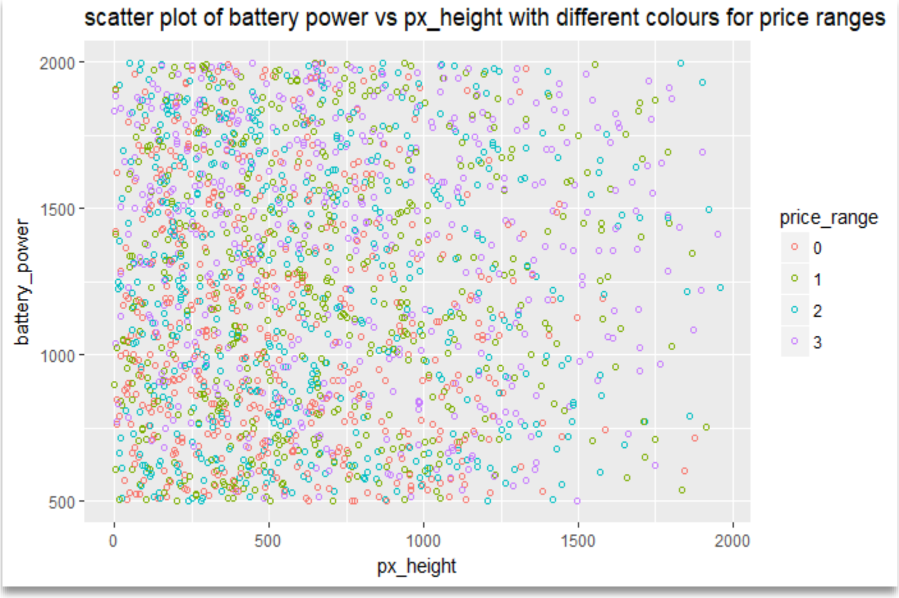
Attribute	Mean	Median	Standard deviation	Minimum value	Maximum value
Battery_power	1238.5	1226	439.418	501	1998
Clock_speed	1.52	1.50	0.816	0.50	3.0
Int_memory	32.05	32.00	18.14	2.00	64.00
Mobile_wt	140.2	141.0	35.39	80.0	200.0
ram	2124	2146	1084.73	256	3998
Px_width	1251.5	1247	432.19	500	1998.00
Talk_time	11.01	11.00	5.46	2.00	20.00

2.4 Visualization

Data visualization helps in understanding the problem of interest very well. In our case, the target variable is multi-variate can be plotted against the independent variables to give more insight into the problem.

Scatter plot can be used to identify the type of relationship between two independent variables. Since our target variable is multi-variate, we can visualize the separation of four classes(with different colors) by plotting independent variables against each other.

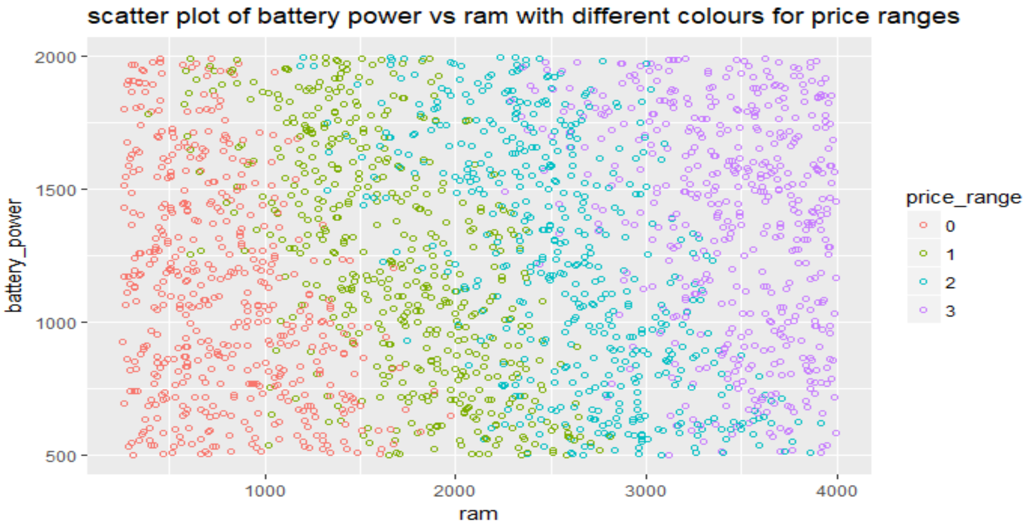
Figure 1 scatter plot of battery power vs px_height



The above scatter plot is obtained by plotting battery_power against px_height(1000 samples) to show how these variables separate the target variable price_range. As we can see from the scatter plot that the relationship is non-linear. Therefore, Non-linearity of the data is one of the problem to address in this objective.

After applying the data preparation techniques, it is found that the variable **ram** is an important feature for deciding the price range of any mobile phones.

Figure 2 Scatter plot of ram vs battery power



One of interesting results found is that ram variable provides a good separation of target variable into four price ranges. Important thing to consider here is how data preparation help to identify essential information in the data.

3.Data Preparation

Data preparation has become an inevitable part of data processing and data mining. [1]It is the process of cleaning and transforming unprocessed data into a form that is easier for extracting information. It involves standardizing, detecting outliers, handling missing values, selecting important features, dimensionality reduction, discovering patterns, finding correlated features and visualizing raw data.

3.1 Missing data and Pre-processing:

The data does not have any missing data, there is no need for handling missing data values in this dataset.

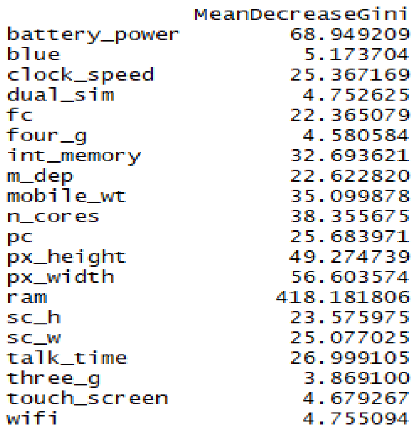
We standardize the quantitative variables to have a zero mean and unit standard deviation. We also create dummies for categorical variables by converting them into factors. We separate the whole data into training and validation data where 60% of data is used as training and 40% is used for validating the models.

3.2 Variable selection

Random forest shows the mean decrease in Gini index for each variable which is an important measure to quantify the importance of variables. After fitting a random forest model using all the variables, we obtain mean decrease in Gini index which is measure of average decrease in Gini index for using the variable during each split in random forest model.

The following figure shows that battery_power and Ram are the important variables contributing majority of decrease in Gini Index while dual-sim, three-g, four-g, touch-screen and wi-fi contribute less.

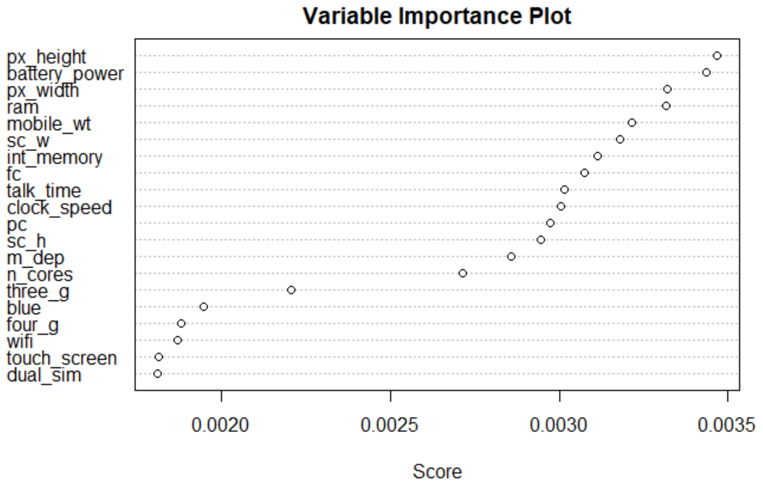
Figure 3 Mean decrease in Gini index for each variable in random forest



```
              MeanDecreaseGini
battery_power        68.949209
blue                  5.173704
clock_speed          25.367169
dual_sim              4.752625
fc                   22.365079
four_g                4.580584
int_memory           32.693621
m_dep                22.622820
mobile_wt            35.099878
n_cores              38.355675
pc                   25.683971
px_height            49.274739
px_width             56.603574
ram                 418.181806
sc_h                 23.575975
sc_w                 25.077025
talk_time            26.999105
three_g               3.869100
touch_screen          4.679267
wifi                  4.755094
```

Multi-class Ada-boosting model is also ideal for obtaining the variable importance. By variable importance here, we mean that improvement in performance of boosting using that particular variable.

Figure 4 Variable importance plot in Ada-boosting



The Above figure shows that battery_power, ram, Px_height, px_width are the important variables contributing majority of improvement in performance while dual-sim, three-g, four-g, touch-screen and wi-fi contribute less.

Based on the above inference about variable importance, We can drop the following variables : dual-sim, three-g, four-g, touch-screen, wifi. We drop these variables because they do not provide any useful information for classifying mobile phones into a price range

4 Methodology & Model comparison

In this section, we explain how we use the prepared data for fitting models like Random forest, Multi-class Ada-boosting and Support vector machine. In the end, we analyze the performance of different models based on different evaluation metrics.

4.2 Random Forest

Random forest is the ensemble tree method which grows number of trees on the boot-strapped samples and uses subset of features for each split to reduce the correlation between the trees.

We fit the Random forest model using important variables obtained in last section and fine tuning is performed to identify the best hyper-parameters(Number of features during each split & number of trees).

4.3 Multi-class Ada-boosting

In multi-class Ada-boosting, the weak learners are grown sequentially where the data is reweighted for every successive weak-learner and all the weak learners are weighted individually to get the final classifier.

Multi Ada-boost model is fitted on important variables and fine tuning is performed to identify the best values for hyper-parameters(Number of iterations & shrinkage parameter).

4.4 Support vector Machine

Support vector machine is a non-probabilistic classifier which assigns new data to either one of the categories of binary variable with the help of support vectors. For Multi-class case, it uses the ‘one-against-one’-approach, in which k(k-1)/2 binary classifiers are trained and the appropriate class is found by a voting scheme.

Since our data is non-linear , we use support vector classifier with kernel trick which gives us a separating hyper-plane by projecting the data into high-dimensional space.

SVM with Radial basis function kernel is fitted and fine tuning is performed to identify the best hyper-parameters(Gamma & Cost parameter).

4.5 Evaluation Results and Model comparisons

We use various evaluation metrics to compare the performance of the models we proposed. They are listed in the following table.

Table 5 Evaluation metrics

Metric	Formula
Accuracy	(TP + TN)/(FP+FN+TP+TN)
Sensitivity or recall	TP/(TP+FN)
Specificity	TN/(FP+TN)
Precision	TP/(TP+FP)
F1-score	(2*recall*precision)/(precision+ recall)

Where, TP – true positives , TN – true negatives, FP- false positives, FN- false negatives.

4.4.1 Results

we present the results obtained for various models that we proposed in the last section. The results are performance of models based on evaluation metrics in the last section. Results shown in the below table is obtained on the validation set that separated initially during pre-processing stage.

We also include the best values obtained for the hyper-parameters of each models along with evaluation metrics discussed in the above table.

Table 6 performance of various models using evaluation metrics

Metric	Random forest				Multi-Ada boost				SVM			
Class	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Accuracy (model)	0.855				0.843				0.92			
Sensitivity or recall	0.92	0.83	0.77	0.86	0.95	0.74	0.73	0.95	0.96	0.92	0.87	0.92
specificity	0.97	0.92	0.92	0.97	0.95	0.94	0.93	0.95	0.99	0.95	0.95	0.99
precision	0.94	0.76	0.73	0.90	0.88	0.81	0.79	0.86	0.94	0.78	0.758	0.92
F-1 measure	0.92	0.81	0.75	0.87	0.91	0.78	0.76	0.90	0.93	0.81	0.76	0.89
Hyper-parameters	Number of features – 10 Number of trees - 700				Number of iterations – 5000 Shrinkage parameter – 1.0				Kernel-RBF(gaussian) Gamma- 0.001 Cost - 100			

4.4.2 Models comparison

As we can see from the above table, SVM with radial basis function outperforms both random forest and Ada-boosting in terms of accuracy, specificity, sensitivity, precision and F-1 score and we choose SVM as best model for classifying mobile phones into their price range.

Since the data is highly non-linear, support vector classifier with non-linear kernel will help us best separate the data by projecting it into high-dimensional space. Performing fine tuning helped us to choose the best non-linear kernel as gaussian with hyper-parameters Gamma-0.001 and cost-100.

Random forest model achieves a accuracy of 85% with hyper-parameters mtry- 10(features during each split) and number of trees – 700.

Ada-boost model achieves a accuracy of 84% after 5000 iterations with shrinkage parameter as 1. Performance of Ada-boost model decreases after 5000 iterations which imply us it is over-fitting the data.

5 Summary & conclusion

In summary, we can say that data preparation techniques like variable selection helps the model to improve their performance in terms of space, time and mis-classification.

In data exploration section, we saw that the data is non-linear as a potential problem to be addressed by the models. As we saw in the results section , SVM with kernel trick performs better when the data is non-linear and models the data very well to help us identify the price range of new mobiles

Random forest performs better when we have both categorical and numerical predictors. Performance of Ada-boost increases as the number of weak learners increases. Both Ada-boost and Random forest can be used for finding the important variables for the objective.

On conclusion, classification of mobile phones using various independent features data can become time consuming without the use of techniques like feature selection and good model selection. I suggest that any entity which is trying to find the price range should well assess the features the mobile phone by feature selection and selected features should be combined with very good model selection.

6 References

[1] https://www.educba.com/data-mining-techniques-2/

[2] https://en.wikipedia.org/wiki/Descriptive_statistics

[3] https://www.talend.com/resources/what-is-data-preparation/

[4]https://www.rdocumentation.org/packages/e1071/versions/1.7-1/topics/svm

[5]https://en.wikipedia.org/wiki/Support-vector_machine

[6]https://www.bing.com/images/search?q=mobiles+with+price&FORM=QBIR