Mobile Price Classification

Abtin Mahyar
Department of Computer Science
Shahid Beheshti University
email:abtinmahyar[at]gmail[dot]com

Abstract

This report gives an overview of the various machine learning algorithms implemented to classify price range of a mobile device based on different features that a mobile has. Exploratory data analysis was performed on the dataset in order to take insights from the data. Feature selection and feature extraction was performed using different machine learning algorithms with the help of Scikit-learn package, and various machine learning models was used to build supervised learning multi-class classifiers, that provided an accuracy of 97% on the test dataset. The dataset was obtained from the popular data science competition portal, Kaggle.

1 Introduction

Making predictions is something a business or system depends on, which is indirectly dependant on data analytics. Data analytics is the science of analyzing raw data in order to make conclusions about that information.This analysis can then be used to optimize processes to increase the overall efficiency of a business or system. The dataset here, that I have obtained from Kaggle, is a sample of various types of mobile devices. A new mobile company wants to compete with other big companies, in order to that, they want to know how to estimate price of mobiles their company creates. This project aims to develop a price range prediction model with comparing different popular models and choose the best one that can recognize the patterns in the data.

2 Methodology

The performed methodology can be described by the following figure, which will be further elaborated on, in the following subsections.

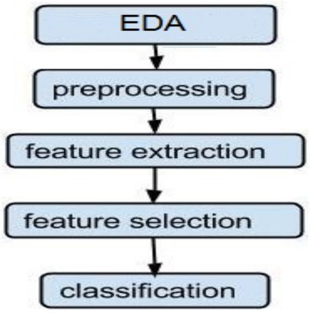

Figure 1: Performed methodology on the dataset

2.1 Exploratory Data Analysis

The data was presented in the form of a csv file on the Kaggle data science competition portal. This dataset has valuable information about the battery power, ram, screen size, number of cores, and other related variables of different mobile devices. This dataset has 2,000 records with various types of fields which each record refers to a particular mobile device. Each record has 20 attributes and a price range which were described completely on the dedicated dataset page.As it can be suspected based on the attributes, some of the features are useless or illegible, which means they must be discarded during the preprocessing section. There is no null values in the whole dataset. Distribution of different features in the dataset is shown in Figure 2.

Correlation analysis between each feature was performed on the data and the results can be seen in Figure 3. As it can be seen most of the features are not correlated with each other, the correlation statistical analysis between features that are probably correlated, according to the above figure, will be performed in the hypothesis subsection.Top correlated features to the target variable (price range) is shown in Table 1.

As it can be seen from Figure 2, none of the distributions has a normal distribution, because

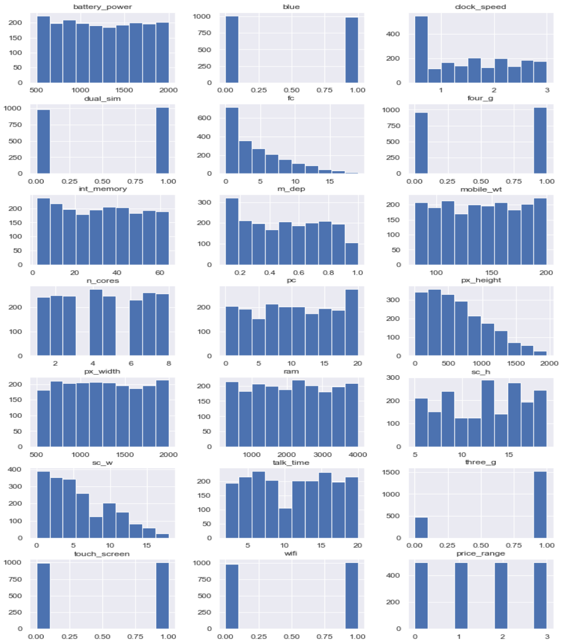

Figure 2: Distribution of each feature in the dataset

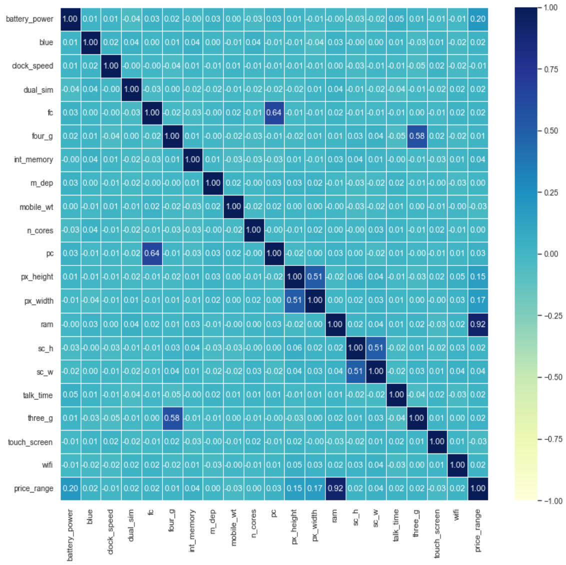

Figure 3: Correlation between each feature in the dataset

most of machine learning algorithms have an assumption that the data should have a normal distribution and have same scale, this issue should be handled in preprocessing section. Also, skewness and kurtosis of different features in the dataset has been calculated, most skewed features are listed in Table 1. Since skewness is a parameter that can badly influence the prediction model, this issue should be handled in preprocessing section.

Feature	Value
Ram	0.91
Battery Power	0.20
Pixel Resolution Width	0.16
Pixel Resolution height	0.14
Internal Memory	0.04

Table 1: Top correlated features to the price range

Feature	Skewness	Kurtosis
Front Camera	1.02	0.28
Pixel Resolution height	0.68	-0.32
Screen Width	0.63	-0.39

Table 2: Most skewed features

2.1.1 Hypothesis testing

2.1.1.1 Normality With using Shapiro-Wilk test, which is a statistical test that assumes that the observations in the sample data are independent and identically distributed, I tested every feature in the dataset to see if they have normal distribution or not, and it comes out that every feature in the dataset does not have Gaussian distribution and their p-values are less than 0.05.

2.1.1.2 Correlation With using Chi-Squared test, which is a statistical which checks whether two categorical variables are correlated or not, I tested every feature in the dataset with each other and I got the same result as Figure 3. Some performed tests and their statistics that have performed are listed as follows:

Feature 1	Feature 2	P-Value	statistic	Result
4G	3G	0.000	679.94	Reject H_0, dependent
4G	Dual Sim	0.92	0.01	Accept H_0, independent

Table 3: Correlation tests using Chi-Squared test

2.1.1.3 Non-parametric Statistical tests Since every feature in the dataset do not have normal distribution, I used non-parametric test in order to compare distributions of the features. With using Mann-Whitney U test, which is a statistical test that checks whether the distributions of two independent samples are equal or not, I tested different features with each other. some performed test and their statistics are listed as below:

Feature 1	Feature 2	P-Value	statistic	Result
Front Camera	Clock Speed	0.00	2605106.0	Reject H_0, different distributions
Battery Power	Pixel Resolution width	0.33	1965063.5	Accept H_0, same distributions

Table 4: Comparing distributions between selected features

2.2 Preprocessing

The preprocessing involved handling skewed data and unscaled features, using normalization method with min-max scaler which transforms the data with following formula:

$$x_{new} = \frac{x-min}{max-min}$$

After normalization still skewness of features from Table 2 are high which can be resulted as a reduction in overall accuracy. In order to reduce influence of these skewed data, some other transformers such as square root, and logarithm are applied to these features. Also, categorical and numerical variables converted to their appropriate data types.

2.3 Feature Extraction

In this step two different extracted from the previous feature based on their correlation to the target variable. First one is pixel resolution of the whole area of the mobile device which is the multiplication of pixel resolution of the width and height of the mobile screen. The second one is the size of the whole screen in centimeters which is the multiplication of size of the width and height of the screen.

2.4 Feature Selection

Feature selection is performed to automatically search for the subset of the attributes in the dataset to find the one with with the highest accuracy. two different method were used in order to calculate most important features for this task which could help the final model to achive better score on the test set. First, a Random Forest classifier trained on the dataset with default hyperparameters and the most important features calculated and stored, in order to fit to the final model. Second, with using forward and backward selection techniques and final model, most important feature for this task are selected and fit to the final model. Both methods selected the same features.

2.5 Classification

Several different classifiers were appplied on the dataset which were generated by carrying out the feature extraction and feature selection phases, but in the following subsections only the top four classifiers with the highest accuracy will be discussed for the evaluation of performance on test data set, These classifiers are listed in Table 5. Also, for evaluating model performances, since the distribution of different classes in the dataset are excatly equal to each other, I used f1 score with macro average method. In addition, k-fold cross validation was performed on the training set before evaluating the test data set, and in this project, for all classifiers, the default k is set to 10. The advanatages of performing k-fold cross validation included that it prevents overfitting of the classifier model and provides generality to the model that could later better classify an independent data set, such as the test data set. The original dataset has split to train and test set, The size of test data set here is 400 instances and train data set is 1600 instances.

Model	Train Accuracy	Test Accuracy
Logistic Regression CV	0.96	0.92
Gradient Boosting	1.0	0.86
Bagging	0.99	0.86
Random Forest	1.0	0.84

Table 5: Top models based on their performances(macro averaged f1)

2.5.1 Logistic Regression CV

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. Multinomial logistic regression can model scenarios where there are more than two possible discrete outcomes. In Logistic Regression CV, model takes two additional hyperparmeters to tune compared to the simple logistic regression, the range of "C" and ratio of L1 constant which is a coefficient when using elastic-net solver. This model tries to find the best values for these hyperparameters. There are also other hyperparameters to tune which were calculated using grid search cross validation which are listed in Table 6. This model achieved 95.81% f1 score on the training set and 93.86% on test set after hyperparameter tunning.

Hyperparameter	Value
C	3792.69
Penalty	L2
solver	lbfgs

Table 6: Tunned hyperparameters for logistic regression classifier

2.5.2 Gradient Boosting

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are in our case decision trees. Like other boosting methods, gradient boosting combines weak "learners" into a single strong learner in an iterative fashion. There are some hyperparameters that has to be tuned for this model which were optimized by the randomized search cross validation. This model achieved 100% f1 score on the training set and 88.30% on test set after hyperparameter tunning.

2.5.3 Bagging

A Bagging classifier is an ensemble meta-estimator that fits base classifiers (which in our case is decision tree) each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. There are some hyperparameters that has to be tuned for this model which were optimized by the randomized search cross validation. This model achieved 94.92% f1 score on the training set and 88.11% on test set after hyperparameter tunning.

2.5.4 Random Forest

Random forest is a spin-off of the decision learning algorithm where many decision trees are created over an arbitrary subspace and the decision at each split of the tree is done by a random process instead of a discrete optimized split, and the mode of the classifications of these individual decision trees forms the final output classification, in our case one of the four classes of price range. There are some hyperparameters that has to be tuned for this model which were optimized by the randomized search cross validation. This model achieved 96.4% f1 score on the training set and 88.62% on test set after hyperparameter tunning.

3 Related Work

There were a number of things that increased accuracy to this project. These include, exploring influence of different scailing method instead of min-max scaler which were used in preprocessing sub section in order to get the higher score. Various types of scaling methods were applied to the dataset and their influences were calculated, the full result is listed in Table 7. As it can be seen from the following table, min-max scaler and maximum absolute scaler has the most influence on the overall result.

Scaler	Train Accuracy	Test Accuracy
Raw data	0.53	0.56
Standard	0.95	0.93
min-max	0.95	0.94
Maximum Absolute	0.95	0.94
Robust	0.95	0.93

Table 7: Influence of different scaling method on the overall accuracy (macro averaged f1)

Another technique that could be applied in the preprocessing section is to apply PCA on the dataset and use different proportion of variance in order to decrease the number of features. It turns out that for high values of proportion of variance (larger than 0.7), this technique does not have much affect on the model's predictions, and model will achieve 56% accuracy on the test set. Also, lower values of proportion of variance can be resulted as a reduction in overall accuracy.

4 Conclusions

As it can be concluded from above sub sections, the best model that can predict the data is Logistic Regression with the hyperparameters which were described in Table 6. After fitting the feature selected dataset to the final model, it's accuracy increased and achieved 96% on training set and 97% on the test set.

References