Mobile Price Range Prediction

Adi Ingrole, Mandar Khatavkar
Data science trainees,
AlmaBetter, Bangalore

Abstract:

To predict "If the mobile with given features will be Economical or Expensive" is the main motive of this research work. Dataset is collected from Alma Better Different feature selection algorithms are used to identify and remove less important and redundant features and have minimum computational complexity. Different classifiers are used to achieve as higher accuracy as possible. Results are compared in terms of highest accuracy achieved and minimum features selected. Conclusion is made on the base of best feature selection algorithm and best classifier for the given dataset. This work can be used in any type of marketing and business to find optimal product (with minimum cost and maximum features). To predict the accuracy of the mobile price range.

1. 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 (e.g.:- 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. The main objective of this project is to build a model which will classify the price range of mobile phones based on the specifications of mobile phones.

Data Description:

Total Rows= 2000
Total features=21

- 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)

2. 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". Machine learning provides us best techniques for artificial intelligence like classification, regression, supervised learning and unsupervised learning and many more 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 today's 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 low, medium, and high or very High

3. Price ranges:

- Very low cost
- Medium cost
- High cost
- Very high cost

4. Steps involved:

- **Exploratory Data Analysis**
  After loading the dataset we performed this method by comparing our target variable that is price_range with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

- **Null values Treatment**
  Our dataset contains a large number of null values which might tend to disturb our accuracy hence we dealt

with Null values with the help of KNN imputer.

- **Handling Discrepancies**
  There were certain discrepancies found in the dataset such as in the screen width feature (sc_w) some values were zero which is impractical in real life so to handle zero values, we replaced them with mean of all available values sc_w for all values of sc_h.

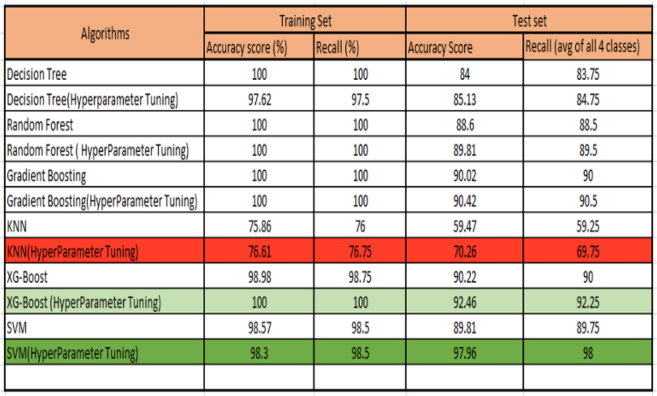
- **Outliers Handling:**
  There were two features with outliers and very few values, one of the best ways to handle outliers is choosing a model which can handle outliers. For other models we removed the outliers as there were few in number using quartiles.

- **Data Preprocessing:**

  This step mainly includes scaling the features for the models that require scaling and splitting the dataset to train and test sets for model evaluation and generating the classification report.

- **Model Selection:**
  We experimented with various models such as
  1. Decision Tree Classifier
  2. SVM
  3. Random Forest Classifier
  4. Gradient Boosting Classifier
  5. XGBoost Classifier
  6. KNN

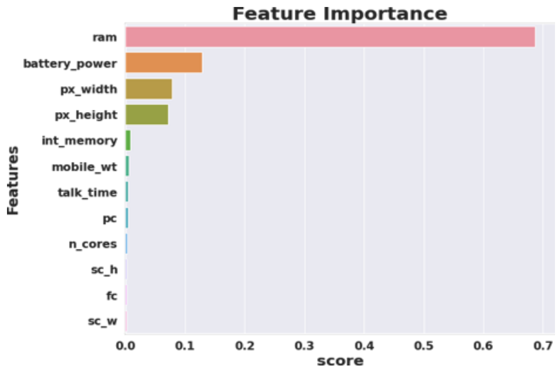
- **Model Performance:**
  We compared 6 classifiers and evaluated them based on overall accuracy & class based accuracy as well.



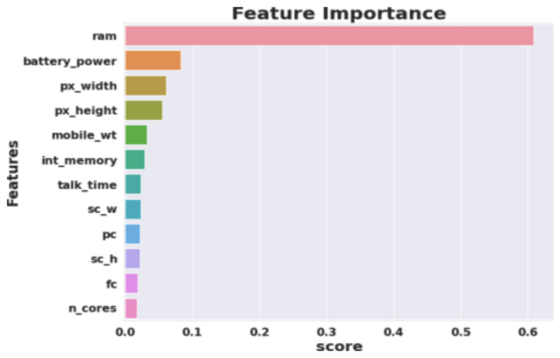
Algorithms	Training Set		Test set	
	Accuracy score (%)	Recall (%)	Accuracy Score	Recall (avg of all 4 classes)
Decision Tree	100	100	84	83.75
Decision Tree(Hyperparameter Tuning)	97.62	97.5	85.13	84.75
Random Forest	100	100	88.6	88.5
Random Forest (HyperParameter Tuning)	100	100	89.81	89.5
Gradient Boosting	100	100	90.02	90
Gradient Boosting(HyperParameter Tuning)	100	100	90.42	90.5
KNN	75.86	76	59.47	59.25
KNN(HyperParameter Tuning)	76.61	76.75	70.26	69.75
XG-Boost	98.98	98.75	90.22	90
XG-Boost (HyperParameter Tuning)	100	100	92.46	92.25
SVM	98.57	98.5	89.81	89.75
SVM(HyperParameter Tuning)	98.3	98.5	97.96	98

  Model performance shows SVM model with best accuracy.

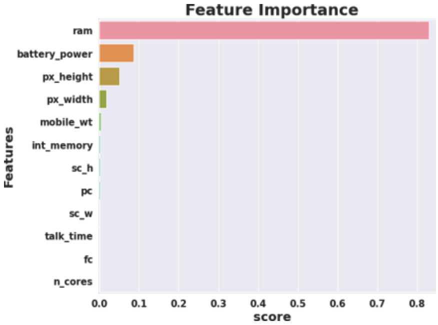
- **Feature Importance:**
  KNN was worst model of these 6 models so we calculated feature importance for only other 5 models.
  1. Decision Tree



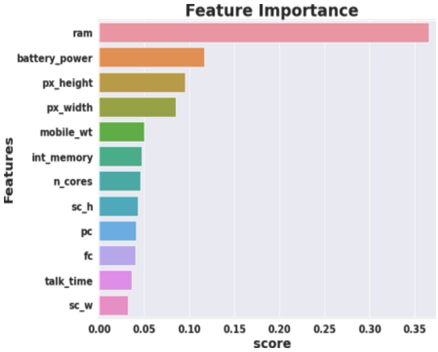
  2. Random Forest



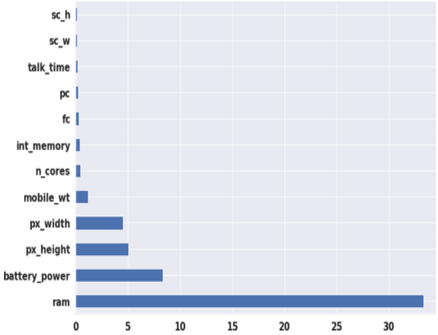
3. Gradient Boost



4. XG boost



5. SVM



RAM, Battery Power, Pixel height and weight contributed the most in predicting the price range.

5. Conclusion:

That's it! We reached the end of our project. We Started with Data understanding, data wrangling, basic EDA where we found the relationships, trends between price range and other independent variables.

We selected the best features for predictive modeling by using K best feature selection method using Chi square statistic.

Implemented various classification algorithms, out of which the SVM (Support vector machine) algorithm gave the best performance after hyper-parameter tuning with 98.3% train accuracy and 97 % test accuracy.

XG boost is the second best good model which gave good performance after hyper-parameter tuning with 100% train accuracy and 92.25% test accuracy score.
KNN gave very worst model performance.
We checked for the feature importances of each model. RAM, Battery Power, Px_height and px_widthcontributed the most while predicting the price range.
In all of these models our accuracy revolves in the range of 70 to 74%.
And there is no such improvement in accuracy score even after hyper parameter tuning.
So the accuracy of our best model is 73% which can be said to be good for this large dataset. This performance could be due to various reasons like: no proper pattern of data, too much data, not enough relevant features.

References-

1. MachineLearningMastery
2. GeeksforGeeks
3. Analytics Vidhya