NEPSI: Is the price range of a mobile phone predictable?

The purpose of this project is to analyze and study the common characteristics of mobile phone products in the market in order to develop a price prediction model. The NEPSI model can help the manufacturer optimize sales planning and guide consumers in purchasing a device that meets their needs.

To achieve this goal, we analyzed a dataset containing 2000 mobile phone products. Each product is described through 21 different attributes, classifying each product into the corresponding price range. These data allowed us to apply supervised machine learning algorithms to create our product.

Exploring the dataset (section: EDA), we were able to verify its integrity and observe how the different price ranges are distributed throughout the dataset. Finally, a correlation matrix allowed us to analyze the mutual relationships. From this inspection, we found that the most correlated features are RAM, battery power, width, and pixel width.

Based on the above analysis and in order to facilitate manufacturers and consumers in evaluating the most relevant parameters for investing in a specific product, we preprocessed (section: Preprocessing) the dataset using Principal Component Analysis (PCA). This way, we were able to focus the dataset on only the features that describe 95% of its variance, reducing the number of attributes required for price prediction. The analysis confirmed that RAM, screen resolution (PIXEL), and battery capacity are relevant attributes for placing a product in a specific price range.

To further validate the choice of predictors, from the aforementioned complete dataset, we decided to use a supervised machine learning algorithm, the Decision Tree model, on the complete dataset (subsection: Overall data). After optimizing the parameters (subsection: Overall data - Parameter tuning), we applied the model, which allowed us to achieve a price prediction accuracy between 80% and 85%, with a True Positive percentage for the four classes ranging from 80% to 95% (subsection: Overall data - train analysis). It is important to emphasize that among the most relevant attributes for the model, the same attributes obtained through PCA were identified for this prediction. Therefore, we were able to engineer and process it accordingly. This optimization led to an increase in accuracy close to 90%, with a True Positive percentage ranging from 90% to 96%.

Subsequently, in order to further refine the classification, we applied ensemble techniques (section: Ensemble), always based on decision tree models. We replicated the sequence of experiments mentioned above, obtaining further confirmation on the most relevant attributes and the higher efficiency of models trained with the engineered dataset rather than the complete dataset. By using a bar chart, we simultaneously evaluated the three models, observing similar performance (RandomForest (RF), Bagging (BG), and AdaBoost (ADA)). The models showed an increased efficiency in predictions compared to the individual decision tree, surpassing the accuracy threshold of 90%.

In addition, we employed a regression model, the Support Vector Machine (SVM) (section: SVM). This model allowed the application of regression techniques on a non-linear dataset, which is not possible with classical linear regression models. The analysis revealed that the SVM model demonstrated the best performance on the entire dataset (SVM\_T), achieving a prediction accuracy of 97%, compared to a slightly lower performance when applied to the engineered dataset (SVM\_P). However, the True Positive percentage was evaluated to be above 90% for all classes.

Despite the possibility of using models that can predict the price using only three characteristics, we decided to use the most performant model (SVM\_T). Nevertheless, there is freedom of choice, based on knowledge and availability, for the inclusion of any indicator. In cases where no indication is given, an average value will be automatically chosen for that characteristic.

In the NEPSI model-Use\_it section, the Nepsi model can be utilized. Initially, a list of features to be considered with their respective average values will be provided as support.

User guide:

Nepsi(<feature>=<value>, <feature>=<value>)

Enter the list, separated by commas (as an example):

<feature>: the indicator of the feature for which a value is to be specified.

<value>: the value for the chosen feature.

USED LIBRARIES:

Numpy, Pandas, Matplotlib, Seaborn, Sklearn

USED ALGORITHMS:

PCA, Decision Tree, Ensemble (Decision tree-based algorithms), SVM

Authors: Guglielmo Tedeschi, Manuel Naviglio, Alessio Franchi