Capstone Project Submission

Instructions:

i) Please fill in all the required information.
ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Name: Sri harish A Email: sriharishanand@gmail.com • KNN imputation of mismatched values • Pie chart for class distribution • Distribution plot for all features • Multi co-linearity plot • Bar chart – Price range vs Numerical features • Count plot of binary categorical variables • Feature Engineering • Function to visualize feature importance of linear based models • Function to predict, fit, transform, and print train, test accuracy scores. • Logistic Regression, Gradient Boosting classifier, K-neighbors classifier • Cross Validation and Hyper parameter Tuning for Logistic regression • Scatter plot for various classes • Box plot for all numerical variables • Outlier detection and imputation • Mean median plot • Line plot for all features • Bar plot- Price range vs Categorical features • Count plot for features • Function to plot AUC and ROC curve • Function to plot feature importance of tree based models • Function for classification report, accuracy score, and confusion matrix. • SVC, Decision tree classifier, Random forest classifier • Cross Validation and Hyper parameter Tuning for SVC • Model Performance Analysis and Selection
Github Link:- https://github.com/Sriharish19/Mobile_price_classification_ml_model

Supervised Machine Learning – Mobile price range prediction

Summary

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 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 dataset provided contains 21 features including the dependent variable which is price range and independent variables such as RAM, battery power, cores, and display size etc. These various features and information can be used to predict the price range of a mobile phone.

The goal of this research is to experiment with machine learning models for anticipating the price range of the mobile phone. Data exploration, feature engineering, choosing an appropriate scoring metric, cross algorithms, cross validation, tuning the algorithms, feature significance analysis and performance evaluation are all part of the research.

We started by importing the data set from Google Drive into our colab notebook, which was in.csv format, and executing simple operations like shape, description and info, is null, and so on to get a basic idea of the dataset's contents.

Data cleaning was the next process to remove unwanted variables and values from our dataset and get rid of any irregularities in it. Since these values disproportionately skew the data and hence adversely affect the results.

This dataset contained few outliers, which might have an impact on the model's efficacy. Outliers were found using the bar plot, and all the rows which contained outliers in them were removed .There were also some mismatched values in some features which were imputed using k-nearest neighbors approach.

Exploratory Data Analysis was performed on the dataset to understand the relationship between the target variable and independent variables. The hypothesis was generated using the EDA. Multivariate analysis and Feature multi co-linearity maps gave insights about the correlation between the independent variables and the impact on dependent variable.

In feature engineering sklearn’s SelectKBest and chi2 were used to select the top 10 important features form the 20 features that was previously available.

Since linear models require standardization .The values were standardized to center them on the mean with a unit standard deviation. Splitting arrays into train and test subsets was accomplished using train test split.

Although the Tree based models are robust to outliers and multi colinearity, it is prone to over fitting. We obtained a 100% accuracy for the train dataset and 89% accuracy for the test dataset. As a result, we intuitively assumed that linear models would be the best fit model for the given dataset which gave an accuracy of 96% in test dataset.

So we performed grid search CV and parameter tuning on Logistic regression and SVC, the Logistic regression model proved to be the best model for the data set, with the best accuracy score of 97 percent. Confusion matrix was used to visualize and summarize the performance of a classification algorithm. Feature importance for the final optimal model showed RAM and battery to be the most important features.