Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Name: Rudrajit Bhattacharyya Email: rudrajitb24@gmail.com Contribution: 1. Importing the Libraries and Loading the Data 2. Data Cleaning/Pre-processing 3. Exploratory Data Analysis a. Distribution of the target variable to check whether the classes are balanced or not b. Distribution of all the independent variables c. How the independent variables vary with different price ranges d. Which features are influential in determining the price ranges e. Correlation analysis 4. Data Preparation for Modeling a. Standardization where required b. Train-Test-Split 5. Building Classification Models a. Logistic Regression b. Random Forest c. Gradient Boosting d. XG Boost e. K Nearest Neighbors f. Support Vector Machines 6. Evaluation of all the Models 7. Comparing all the Models 8. Conclusion
Please paste the GitHub Repo link.
Github Link:- https://github.com/Rudrajit12/SupervisedML_Classification_Capstone_Project
Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

In the competitive mobile phone market companies want to understand sales data of mobile phones and factors which drive the prices. Mobile phones come in all sorts of prices, features, specifications etc and estimating the price of mobile phones is an important part of consumer strategy. Deciding on the correct price of a product is very important for the market success of a product. The objective of this work is to find out some relation between features of a mobile phone and its selling price. In this problem, we didn't need to predict the actual price but a price range indicating how high the price is.

The dataset contains 2000 records of mobile phone information with 21 features which were a mix of categorical and numerical values. The dataset was almost a cleaned one with no null values present or duplicate records found. Few records had to be dropped from the dataset as those contained zero values which are unreal for pixel resolution and screen width.

With the help of EDA, we got to know that the target classes are mostly balanced and do not contain much difference. Most of the categorical features had a similar distribution except 'three_g' where there were very few records of mobile phones not having 3G access. Most of the numerical features follow an uniform distribution except a few features which were right skewed. The categorical features do not vary much with different price ranges but we noticed a slight increase in the count for each feature for the very high cost category. RAM has the strongest correlation with the target variable and thus it is the most influential factor in determining the price ranges. Battery power and pixel resolution are slightly correlated with the target variable which means these four features are influential in determining the price ranges. Apart from these features, no other feature had a good correlation with the target variable.

Six classification models were implemented and metrics were calculated where we observed that all the models had done a fair job in predicting the price ranges **(Logistic Regression, Random Forest, Gradient Boosting, XGBoost, KNN, SVM)**. Logistic Regression and SVM has performed the best out of the six classification models by achieving an accuracy of 96% whereas the other models achieved an accuracy near 90%. Metrics such as accuracy, precision, recall and roc auc score was calculated to compare all the models where Logistic Regression and SVM has performed the best in all the metrics used. We found out the feature importances through the tree based methods and used them to remodel the data again which has shown a bit of improvement in few of the algorithms. Generally, tree based methods have performed poorly in comparison to the other methods.

Please paste the drive link to your deliverables folder. Ensure that this folder consists of the project Colab notebook, project presentation and video.

Drive Link:
https://drive.google.com/drive/folders/1MHRC6UzZkpQWOp3QLRPfuC-LX7_yX1UQ?usp=sharing