Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
1. Puroshotam Kumar Singh: purshotamsingh61@gmail.com • Exploratory data analysis – univariate and multivariate analysis. • Data Wrangling – checking missing values, outliers, features modification. • Fitting Models – splitting the data, applying algorithms, evaluating, model explanation. • Presentation, Technical documentation. 2. Vikram Pandey: pandey.vicky@yahoo.com • Exploratory data analysis – univariate and multivariate analysis. • Data Wrangling – checking missing values, outliers, features modification. • Fitting Models – splitting the data, applying algorithms, evaluating, model explanation. • Presentation, Technical documentation.
Please paste the GitHub Repo link.
Puroshotam's Github Link:- https://github.com/PuroshotamSingh/Mobile-Price-Range-Prediction Vikram's Github Link:- https://github.com/vickypandey07/Mobile-Price-Range-Prediction
Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)
Mobile phones have become a necessity for every individual nowadays. People want more features and best specifications in a phone and that too at cheaper prices. In the competitive mobile phone market companies want to understand sales data of mobile phones and factors which drive the prices. The data provided had features such as battery power, ram, internal memory, mobile weight, camera resolutions, pixel dimensions, screen dimensions, 3g, 4g, touchscreen, dual sim, talktime, wifi, bluetooth and target variable price range. The problem statement was to build a machine learning model that could predict the price range values, given other variables. The first step involved exploratory data analysis, where we tried to identify patterns, trends, correlation, relationships of each variable with our dependent variable. We tried to figure out impact of each variable in determining the price range values. The second step involved data wrangling in which we tried to check data integrity,

missing values, outliers and performed feature modifications.

Next step involved implementing machine learning algorithms on our splitted and standardized data and evaluate the performance using several evaluation metrics. Two algorithms were used namely; Random Forest classifier and XGBoost classifier. The best performance was given by the XGBoost model. We also implemented hyperparameter tuning to improve model performance.

Finally, we implemented SHAP technique to understand the working of the model. We saw ram, battery power, pixel height and pixel width were the top contributors in determining price ranges. Higher the values of these, led to higher price range.

The accuracy of our best model was 0.89 and 0.85 for training and test set respectively. Although, the difference is still 4, considering the simplicity and less no. of observations, this can be considered a good model. Performance can be improved even further by applying fine tunings and gathering a greater number of observations so that the models can identify more patterns and become less prone to overfitting. With evolution of new technology, these numbers can change in future hence there will always be a need to check on the model from time to time.