Mobile Price Range Prediction

Name:	**Abhishek Prasad**
Registration No./Roll No.:	20011
Institute/University Name:	IISER Bhopal
Program/Stream:	EECS
Date of Submission:	November 17, 2023

1 Introduction

This project aims to forecast the price range of mobile phones based on diverse specifications. The pricing is categorized into four groups: inexpensive, moderate, economical, and costly, denoted by 0, 1, 2, and 3, respectively Using the attributes shown in Figure 1, we wish to estimate the price range for the mobile devices. The training has data of 2000 mobile devices. The dataset is well-balanced across all classes and contains no null values..

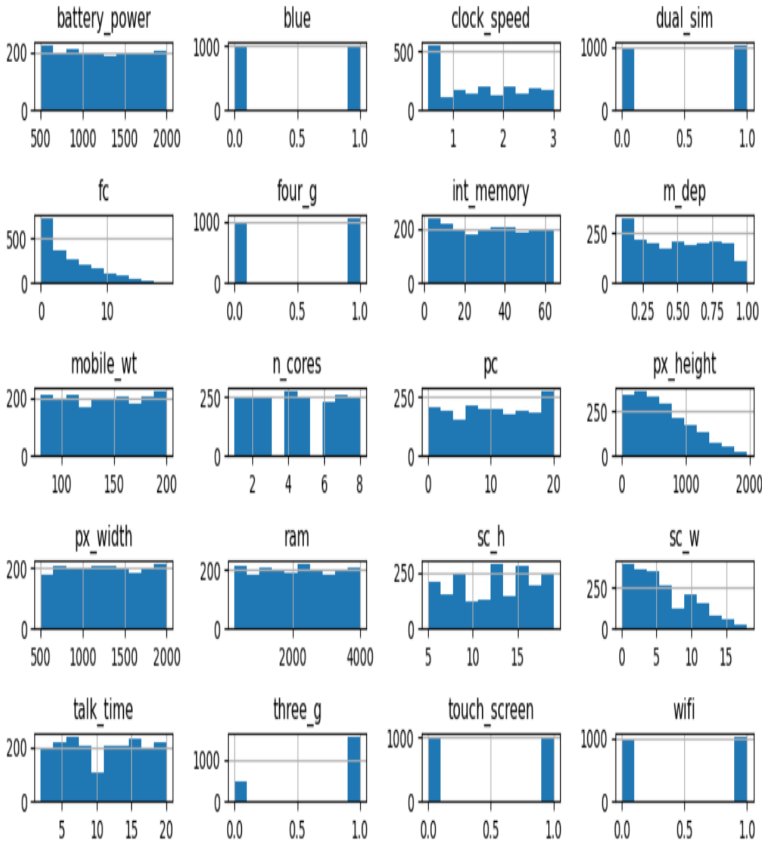

Figure 1: Features Overview

2 Methods

2.1 One hot encoding

We did one hot encoding of the following coloumns:'wifi', 'blue', 'dual_sim', 'three_g', 'touch_screen', 'four_g'. Below figure shows the correlation matrix

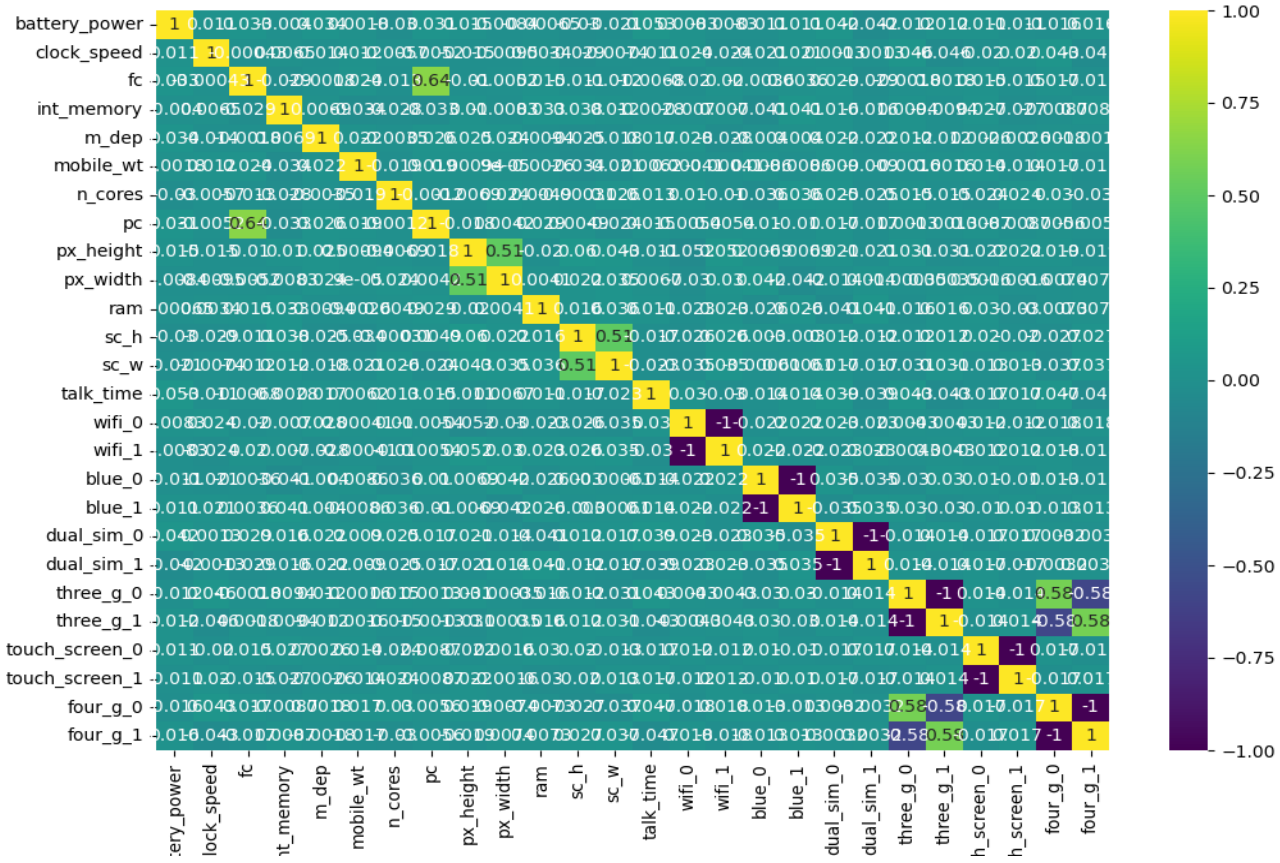

Figure 2: Correlation matrix

2.2 Train-Test Split

The dataset is randomly split into training and testing sets by a factor of 0.7, i.e., 70% data (1400 rows) is selected for training and 30% data (600 rows) is selected for testing the model.

2.3 Models Used For Classification

This problem is clearly a classification problem as we have four discrete classes (0, 1, 2 and 3). Hence, the main classification algorithms used for this problem are:

- Logistic Regression
- K-Nearest Neighbors
- Gradient Boosting
- Random Forest Classifier
- Support Vector Classifier

We then ran grid search cv on all this classifier and got baseline performance.The complete project can be found here.

2.4 Feature selection

We did feature selection to enhance model efficiency and interpret-ability in a classification task. Utilizing the chi-squared test, I explore various classifiers and feature counts to pinpoint the combination delivering the highest average F1 macro score via cross-validation. The dataset is divided into training and testing sets, and class statistics are printed, The primary objective is to identify a subset of informative features. Selected features are 'battery_power', 'int_memory', 'px_height', 'px_width', 'ram'. With this features we did the gridsearchcv and found the hyperparameter for all the classifier

Table 1: Hyper-parameters of different models

Models	Best Hyper-parameter features
K-NN	• n-neighbours: [15] • weights : ['distance'], • 'metric' : ['manhattan']
Random forest	• n estimators: [200] • max depth: [200]
Gradient Boosting	• $n_estimators$: [200], $learning_rate$: [0.2]
Logistic Regression	• C:[0.001] •solver:['newton-cg'] •penalty:['l2']
SVM	• C :[1.189207115002721] • gamma: [1] • kernel:['linear']

3 Analysis of Results

Table 2 shows the recall, precision, accuracy and f measure for all the classification models used in this project

Table 2: Performance Of Different Classifiers Using All Terms

Classifier	Precision	Recall	Accuracy	F-measure
K-NN	0.94761	0.94765	0.94833	0.94754
Random forest	0.9001	0.8992	0.9025	0.8991
Gradient Boosting	0.8995	0.9004	0.9025	0.8999
Logistic Regression	0.9842	0.9851	0.985	0.98447
SVM	0.9745	0.9754	0.975	0.9746

4 Discussions and Conclusion

I found out that Logistic Regression provides the best f-measure as well accuracy and hence I'll use it to build my model. This model only use 5 feature out of 20 given feature making it a very efficient model.

5 References

- Lecture notes by Dr.Tanmay Basu.
- Machine Learning by Tom M. Mitchell.
- `https://www.kaggle.com`

- https://scikit-learn.org
- https://towardsdatascience.com

References