安徽理工大學

ANHUI UNIVERSITY OF SCIENCE & TECHNOLOGY

计算机科学与工程学院

Python Data Analysis Literature Reading Report

学　　号　2021290010

姓　　名　Mahadi Sajjad Neloy

指导教师　Gaoming Yang

完成时间　07-09-2022

Title: Data Analysis of Mobile Phone Prices

Abstract：

Data analysis has become one of the most important aspects in modern times. Industries , businesses, and schools are using data analysis to achieve better results. With the help of Python, which consists of numpy , pandas, etc., data has been well analyzed and the results have become better and better every day given the datasets given. Using a dataset provided by kaggle of mobile phones, we have managed to achieve better accuracy on price estimation depending on the features of every phone. Big companies like Apple, Samsung, etc. are involved in this analysis and their results are analyzed. Using machine learning, the results achieved are 90 % precision.

keyword:

Mobile phones, cell phones, wireless communications, mobile healthcare, data analysis.

1. Introduction：

The mobile industry is a division of the telecommunications sector that focuses on mobile phones, phone service, and peripheral equipment. The global number of smartphone subscribers has surpassed six billion and is anticipated to rise by several hundred million in the next few years. The nations with the most smartphone users are China, India, and the United States. In terms of market size and models, this sector has been consistently rising and growing. This massive number of sales generates massive amounts of data, data that can be studied and used to predict sales patterns, determine justifiable price brackets, and make meaningful business decisions, improve yearly supply, implement robust marketing strategies, improve user service, and ultimately grow the business. Mobile phones are cellular phones that provide basic features such as text messaging, voice calling, audio and video visualization, and a camera. Smart phones are cellular phones that have advanced computing capabilities such as Wi-Fi, web browsing, third-party applications, and mobile payment; solutions for information management such as documents, emails, and contacts; inbuilt GPS applications; and features such as voice and video calls and web access. The purpose of this project is to predict the degree of price range from the "Mobile Price Range Prediction" data set using EDA approaches, using multiple classification models that will offer an insight into our data and be able to estimate the price bracket for new models. In addition, we will answer some questions and display a few patterns utilizing these.

2. relation work

For most epidemiological analyses and modelling work, an estimate of the population residing in a specific region at a particular time provides an estimate of the denominator, which is the number of unique individuals that spend most of their time in a given area(1-5). Many researchers have been analyzing mobile phone data for different perspective(3-7). All their analysis have a lot to show in terms of mobile usage, network usage and only a fill have dive deep into price ranging. In this work , data of different phone companies is analysis and all the results are shown completely.

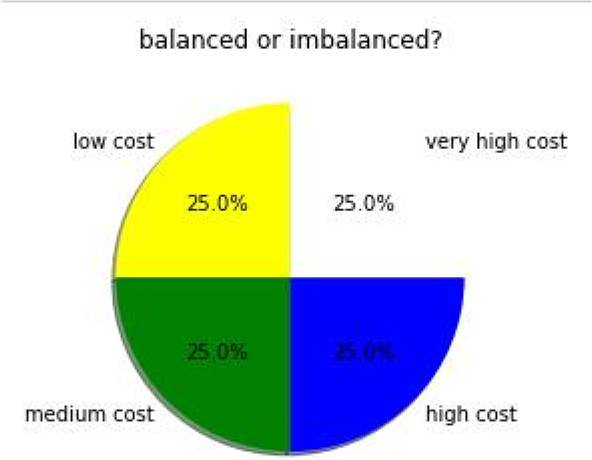
3. Methods

Using the "Mobile Price Range Prediction Data Set" for this Classification project is best due the the fact that this data set has around "2000" observations/listings in it with 21 columns in it's raw form. This data set is a mix between categorical, numeric and Ordinal values. These observations include features such as mobile phone battery capacity, whether the phone has Bluetooth or not, various information on RAM size, internal memory size, camera of a specific model, connectivity features, pixel and screen size, and more.

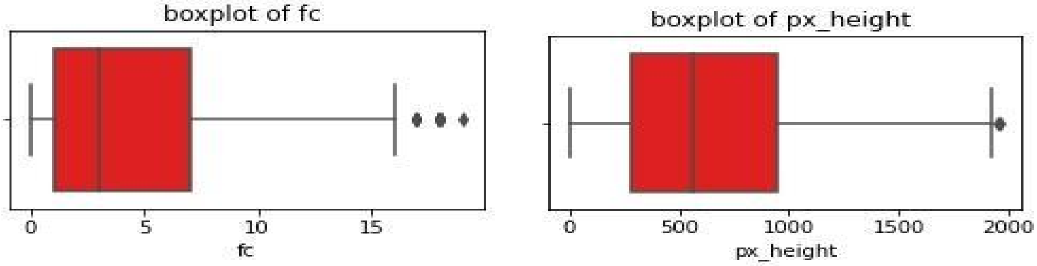
A) Importing necessary libraries, Loading and understanding data. Importing required packages for data manipulation.(Numpy, Pandas, matplotlib and seaborne)
B) Finding shape of the data set to get number of rows and columns. This data set have "2000" columns and '21' rows.(From unmodified/raw data)
C) Analyzing some basic information of data set (using .info function). There are three different data types (int64, object, float64) occupying around '328.2 KB' of memory. encounter "0" Null values is . Though it's highly unlikely to encounter such data sets. Here it works in our favour.
D) Next, Checked for duplicate records/observations for termination. But surprisingly, this data set had no duplicate observations.
E) Checked for number of unique values for each column in order to identify categorical features of this data set.
F) Using describe function to check the min/max values, mean, standard deviation and spread of numeric columns.
G) Now, we are done with cleaning, and our data is now ready for analysis.

Exploratory Data Analysis

A) Exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. We began our EDA with univariate analysis. We check if our data is balanced or imbalanced. This will help us on determining the method to approach this problem. In this case our data is balanced.

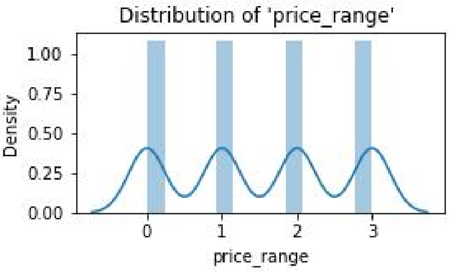


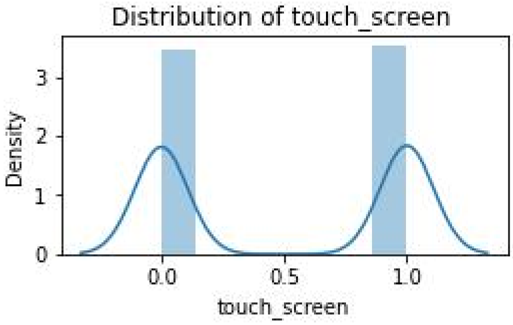
B) Where checks have been performed for type of distribution, skewness and outliers. Considering outliers, we encountered outliers in few features like “Front Camera megapixels” and “pixel height”. Our dataset is fairly clean from outliers. We have very few of them which won't disturb our analysis.

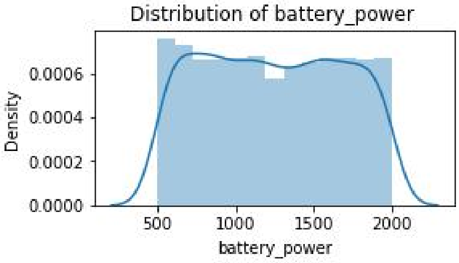


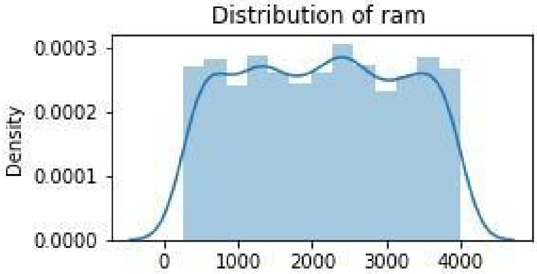
C) We will be keeping them in our analysis and check how models fits. Reason we are considering outliers is because few important variables like "Front Camera megapixels" have a maximum value of about 20 Megapixels from dataset. But, mobiles in market today do have camera having 20+ Mp. It might be that those are legit observations.

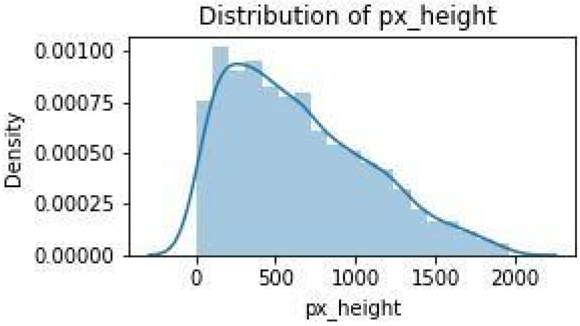
Analyzing distributions of variables.

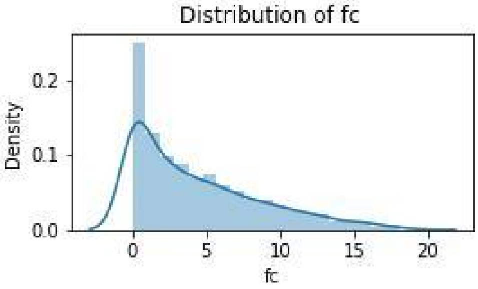




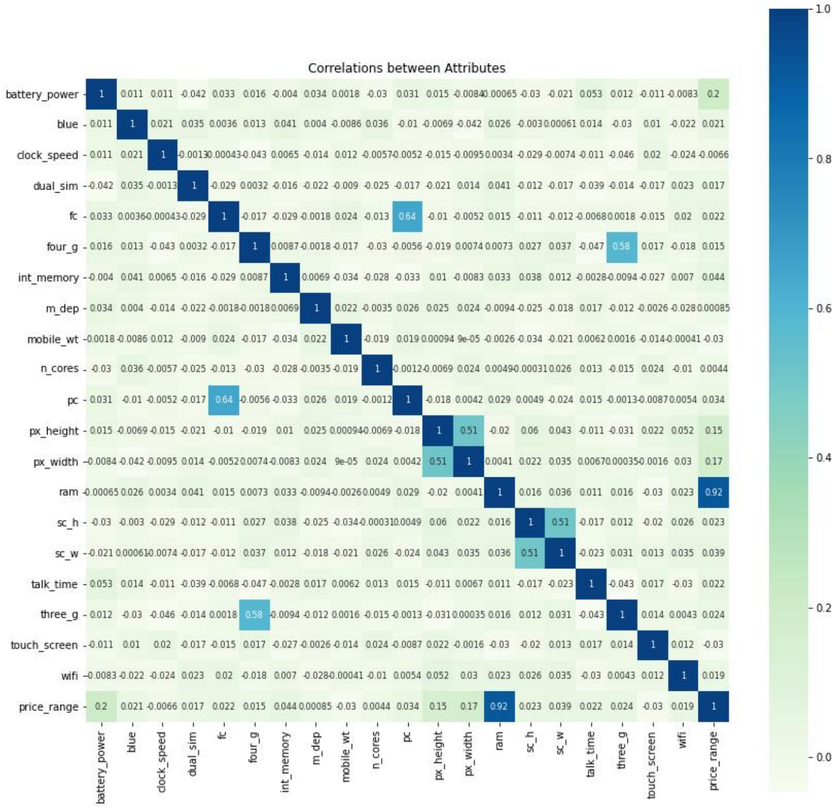






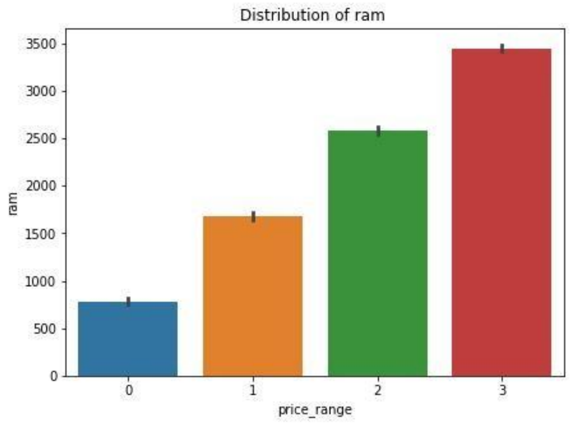


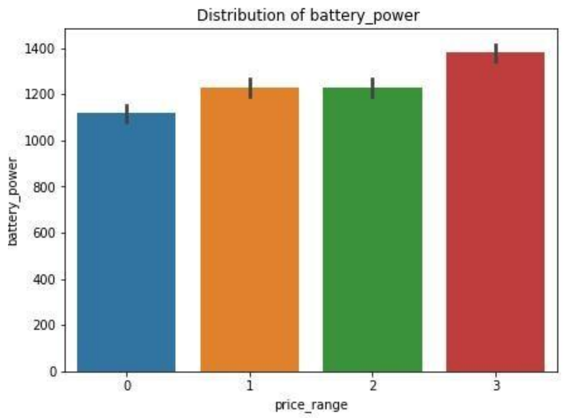
D) Most of the features have uniform distribution including our target variable and can be clearly seen using heatmap. But, we have some skewed variables like “Front camera MP” and “Pixel Height”. It’s difficult to choose a algorithm in this case, but as it’s a multi-class classifier we’ll stick with KNN or Tree based models.



	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	pc	px_height	px_width	ram	sc_h	sc_w	talk_time	three_g	touch_screen	wifi	price_range
battery_power	1	0.011	0.011	-0.042	0.033	0.016	-0.004	0.034	0.0018	-0.03	0.031	0.015	-0.0084	-0.00065	-0.03	-0.021	0.053	0.012	-0.011	-0.0083	0.2
blue	0.011	1	0.021	0.035	0.0036	0.013	0.041	0.004	-0.0086	0.036	-0.01	-0.0069	-0.042	0.026	-0.003	0.00061	0.014	-0.03	0.01	-0.022	0.021
clock_speed	0.011	0.021	1	-0.0013	-0.00043	-0.043	0.0065	-0.014	0.012	-0.0057	-0.0052	-0.015	-0.0095	0.0034	-0.029	-0.0074	-0.011	-0.046	0.02	-0.024	-0.0066
dual_sim	-0.042	0.035	-0.0013	1	-0.029	0.0032	-0.016	-0.022	-0.009	-0.025	-0.017	-0.021	0.014	0.041	-0.012	-0.017	-0.039	-0.014	-0.017	0.023	0.017
fc	0.033	0.0036	-0.00043	-0.029	1	-0.017	-0.029	-0.0018	0.024	-0.013	0.64	-0.01	-0.0052	0.015	-0.011	-0.012	-0.0068	0.0018	-0.015	0.02	0.022
four_g	0.016	0.013	-0.043	0.0032	-0.017	1	0.0087	-0.0018	-0.017	-0.03	-0.0056	-0.019	0.0074	0.0073	0.027	0.037	-0.047	0.58	0.017	-0.018	0.015
int_memory	-0.004	0.041	0.0065	-0.016	-0.029	0.0087	1	0.0069	-0.034	-0.028	-0.033	0.01	-0.0083	0.033	0.038	0.012	-0.0028	-0.0094	-0.027	0.007	0.044
m_dep	0.034	0.004	-0.014	-0.022	-0.0018	-0.0018	0.0069	1	0.022	-0.0035	0.026	0.025	0.024	-0.0094	-0.025	-0.018	0.017	-0.012	-0.0026	-0.028	0.00085
mobile_wt	0.0018	-0.0086	0.012	-0.009	0.024	-0.017	-0.034	0.022	1	-0.019	0.019	0.00094	9e-05	-0.0026	-0.034	-0.021	0.0062	0.0016	-0.014	-0.00041	-0.03
n_cores	-0.03	0.036	-0.0057	-0.025	-0.013	-0.03	-0.028	-0.0035	-0.019	1	-0.0012	-0.0069	0.024	0.0049	-0.00031	0.026	0.013	-0.015	0.024	-0.01	0.0044
pc	0.031	-0.01	-0.0052	-0.017	0.64	-0.0056	-0.033	0.026	0.019	-0.0012	1	-0.018	0.0042	0.029	0.0049	-0.024	0.015	-0.0013	-0.0087	0.0054	0.034
px_height	0.015	-0.0069	-0.015	-0.021	-0.01	-0.019	0.01	0.025	0.00094	-0.0069	-0.018	1	0.51	-0.02	0.06	0.043	-0.011	-0.031	0.022	0.052	0.15
px_width	-0.0084	-0.042	-0.0095	0.014	-0.0052	0.0074	-0.0083	0.024	9e-05	0.024	0.0042	0.51	1	0.0041	0.022	0.035	0.0067	0.00035	-0.0016	0.03	0.17
ram	-0.00065	0.026	0.0034	0.041	0.015	0.0073	0.033	-0.0094	-0.0026	0.0049	0.029	-0.02	0.0041	1	0.016	0.036	0.011	0.016	-0.03	0.023	0.92
sc_h	-0.03	-0.003	-0.029	-0.012	-0.011	0.027	0.038	-0.025	-0.034	-0.00031	0.0049	0.06	0.022	0.016	1	0.51	-0.017	0.012	-0.02	0.026	0.023
sc_w	-0.021	0.00061	-0.0074	-0.017	-0.012	0.037	0.012	-0.018	-0.021	0.026	-0.024	0.043	0.035	0.036	0.51	1	-0.023	0.031	0.013	0.035	0.039
talk_time	0.053	0.014	-0.011	-0.039	-0.0068	-0.047	-0.0028	0.017	0.0062	0.013	0.015	-0.011	0.0067	0.011	-0.017	-0.023	1	-0.043	0.017	-0.03	0.022
three_g	0.012	-0.03	-0.046	-0.014	0.0018	0.58	-0.0094	-0.012	0.0016	-0.015	-0.0013	-0.031	0.00035	0.016	0.012	0.031	-0.043	1	0.014	0.0043	0.024
touch_screen	-0.011	0.01	0.02	-0.017	-0.015	0.017	-0.027	-0.0026	-0.014	0.024	-0.0087	0.022	-0.0016	-0.03	-0.02	0.013	0.017	0.014	1	0.012	-0.03
wifi	-0.0083	-0.022	-0.024	0.023	0.02	-0.018	0.007	-0.028	-0.00041	-0.01	0.0054	0.052	0.03	0.023	0.026	0.035	-0.03	0.0043	0.012	1	0.019
price_range	0.2	0.021	-0.0066	0.017	0.022	0.015	0.044	0.00085	-0.03	0.0044	0.034	0.15	0.17	0.92	0.023	0.039	0.022	0.024	-0.03	0.019	1

E) Bi-variate analysis. Checking for relationship between dependent and independent variables.





Every feature was uniformly distributed considering our target variable. Only "RAM" and "Battery power" showed some amount of relation to our target variable. Here we can estimate that these two features will mostly affect our target variable.

Preparation For Prediction Model

A) Feature engineering

Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning. In order to make machine learning work well on new tasks, it might be necessary to design and train better features. Feature engineering facilitates the machine learning process and increases the predictive power of machine learning algorithms by creating features from raw data.

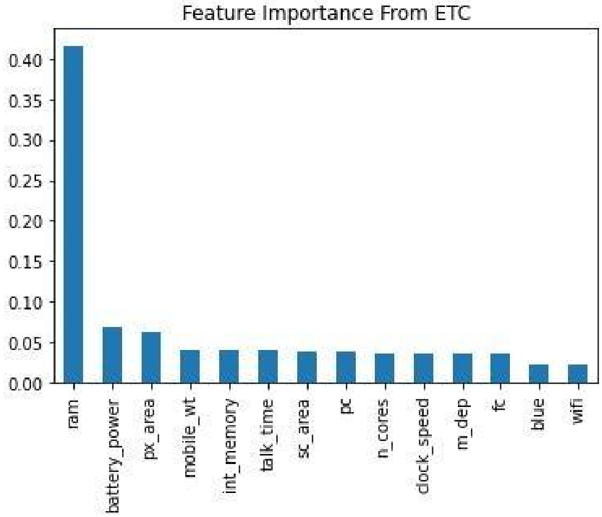
B) New Features created were,

We introduced new feature "px_area" which will be combination of "px_width" and "px_height". We multiplied these two to represent area of particular pixel.

Also, "sc_area" was introduced, taking place of "sc_h" and "sc_w". Here again we multiplied these two to get screen area.

C) Extra Tree Classifier for feature selection.

Extremely Randomized Trees Classifier(Extra Trees Classifier) is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a “forest” to output it’s classification result. In concept, it is very similar to a Random Forest Classifier and only differs from it in the manner of construction of the decision trees in the forest.



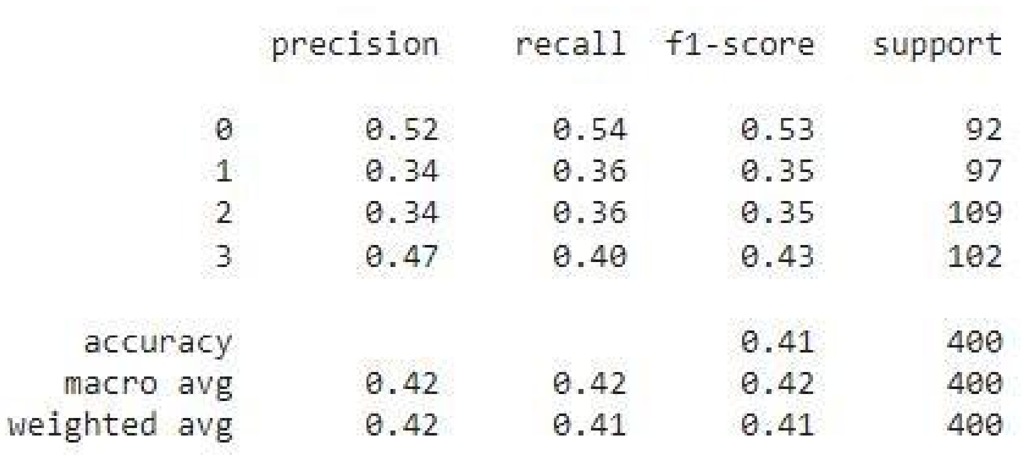
As we can analyze from above, it’s difficult to choose features from this test. Therefore we will be using all features during classification.

Implementing classification models

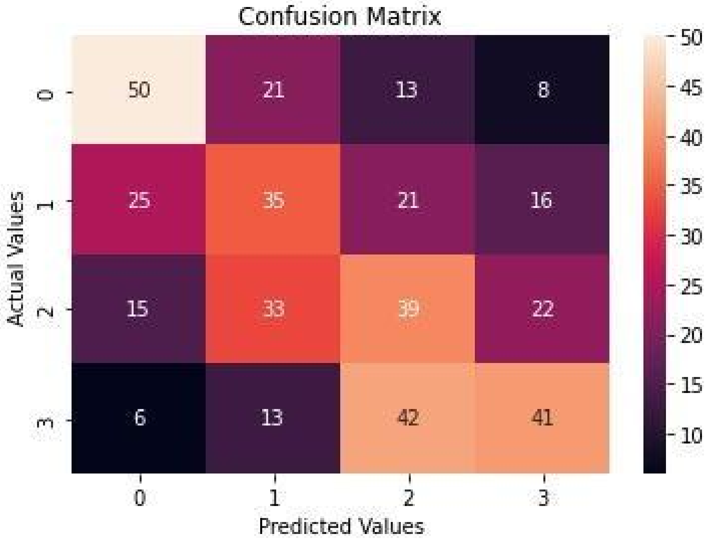
A) Implementing KNN Classifier

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. It’s a non-parametric algorithm, which means it does not make any assumption on underlying data. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

Evaluation Metrics for KNN Classifier



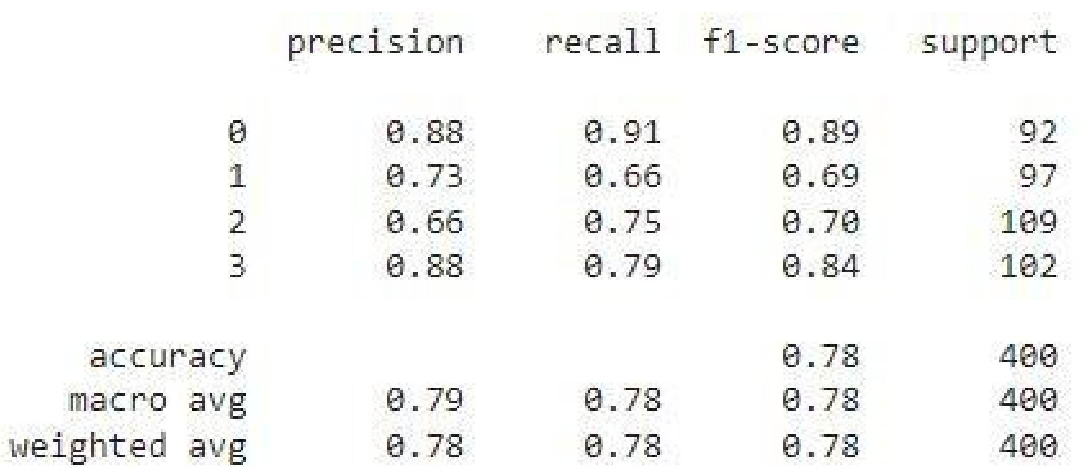
```
              precision    recall  f1-score   support

           0       0.52      0.54      0.53        92
           1       0.34      0.36      0.35        97
           2       0.34      0.36      0.35       109
           3       0.47      0.40      0.43       102

    accuracy                           0.41       400
   macro avg       0.42      0.42      0.42       400
weighted avg       0.42      0.41      0.41       400
```



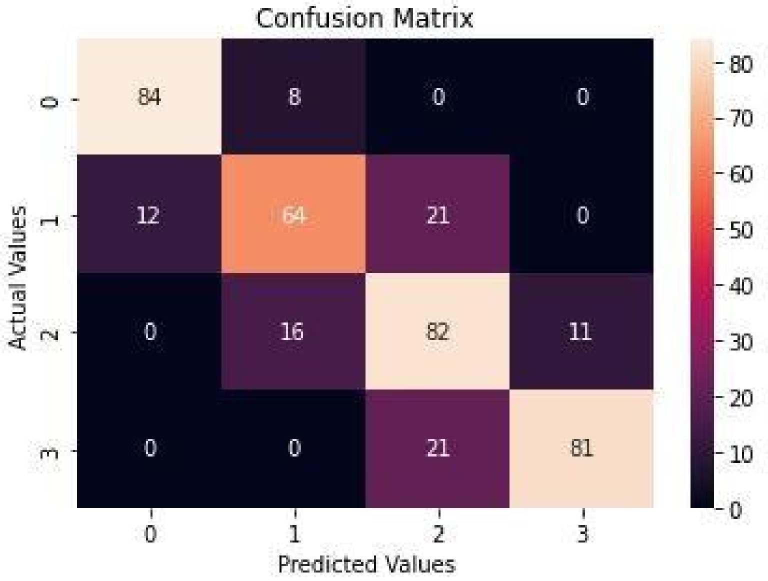
B) Implementing Decision Tree Classifier

Decision Tree Classifier is a structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node.

Evaluation Metrics for Decision Tree Classifier



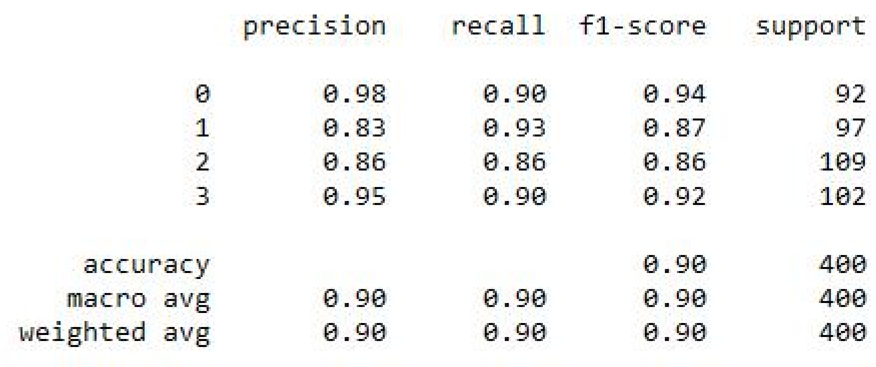
```
              precision    recall  f1-score   support

           0       0.88      0.91      0.89        92
           1       0.73      0.66      0.69        97
           2       0.66      0.75      0.70       109
           3       0.88      0.79      0.84       102

    accuracy                           0.78       400
   macro avg       0.79      0.78      0.78       400
weighted avg       0.78      0.78      0.78       400
```



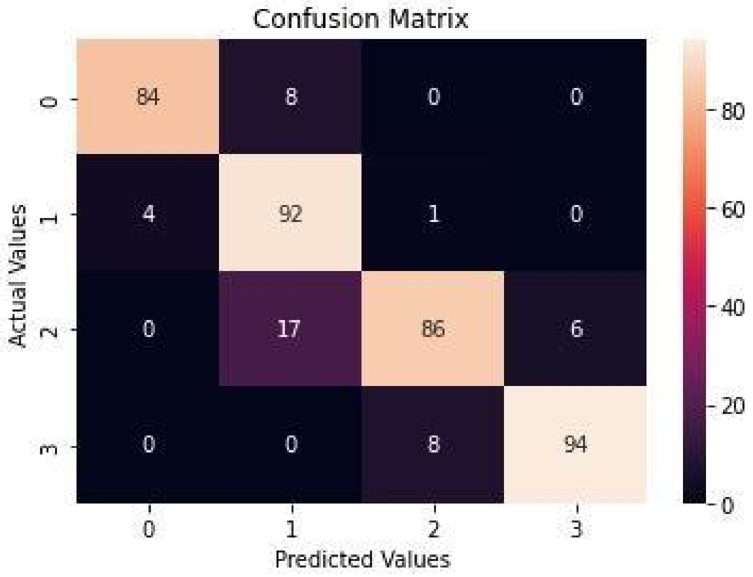
C) Implementing XGBoost Classifier

The XGBoost algorithm is effective for a wide range of regression and classification predictive modeling problems. It is an efficient implementation of the stochastic gradient boosting algorithm and offers a range of hyperparameters that give fine-grained control over the model training procedure.

Evaluation Metrics for XGBoost Classifier



```
              precision    recall  f1-score   support

           0       0.98      0.90      0.94        92
           1       0.83      0.93      0.87        97
           2       0.86      0.86      0.86       109
           3       0.95      0.90      0.92       102

    accuracy                           0.90       400
   macro avg       0.90      0.90      0.90       400
weighted avg       0.90      0.90      0.90       400
```

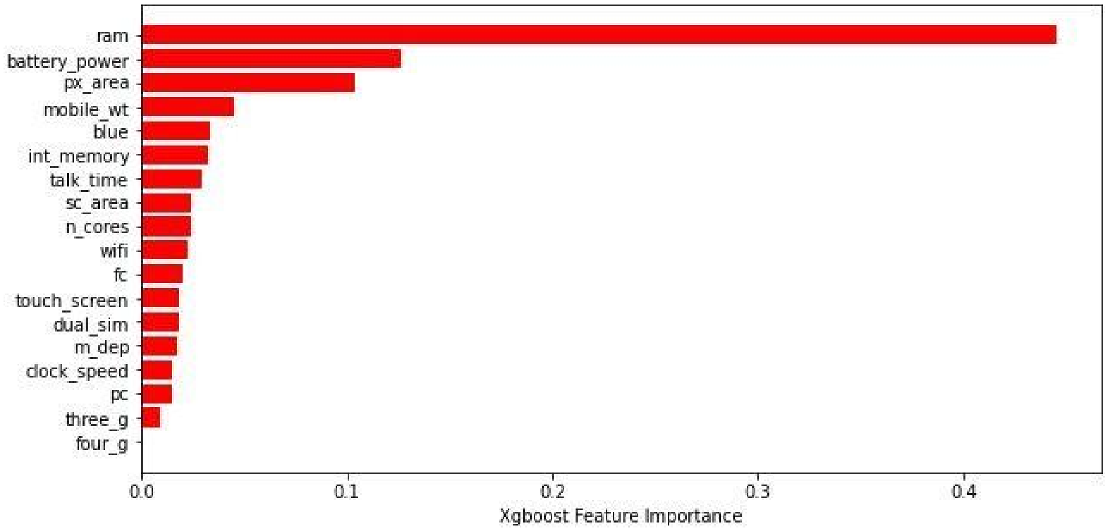


Evaluation metrics Comparison table

	KNN	Dec. T	XG
Accuracy	0.42	0.78	0.90

D) Actual Important Features by XGB Classifier

Feature Importance refers to techniques that calculate a score for all the input features for a given model. The scores simply represent the “importance” of each feature. A higher score means that the specific feature will have a larger effect on the model that is being used to predict a certain variable. Below we have a bar-plot representing feature importance in sorted order.



As we can see here, our top features which were important for predicting target variable are “ram”, “battery power”, “pixel area” and “mobile weight”.

Conclusion

Considering multi class classification, Logistic Classification won’t fit this dataset well enough. Therefore, we used KNN, Decision Tree and XGBoost Classifier here. Considering outliers, our dataset is fairly clean. We have very few of them which won't disturb our analysis. We encountered outliers in features like “Front Camera megapixels” and “pixel height”.While implementing KNN Classifier, even using best parameters and cross validation we observed, (accuracy score was - 0.41, Precision 0.52, 0.34, 0.34, 0.47 , Recall - 0.54, 0.36, 0.36, 0.40 , F1 Score - 0.53, 0.35, 0.35, 0.43) which is really not acceptable. And thus we will have to reject this model.
While implementing Decision Tree Classifier, tuning it’s hyper parameters and using cross validation we observed metrics were way better than KNN. Here the model fit’s well to test data and can be used to predict actual price range (accuracy score was - 0.78 , Precision - 0.88, 0.73, 0.66, 0.88 , Recall - 0.91, 0.66, 0.75, 0.79 , F1 Score - 0.89, 0.69, 0.70, 0.84). Implementing XGBoost Classifier, this is where we observed the highest values of all metrics compared to Decision Tree and KNN (accuracy score was - 0.89, Precision - 0.95, 0.79, 0.91, 0.94 , Recall - 0.91, 0.95, 0.79, 0.92 , F1 Score - 0.92, 0.86, 0.84, 0.93) .

Reference：

1. Kishore, Nishant, et al. "Measuring mobility to monitor travel and physical distancing interventions: a common framework for mobile phone data analysis." *The Lancet Digital Health* 2.11 (2020): e622-e628.
2. Cinque, Marcello, et al. "How do mobile phones fail? a failure data analysis of symbian os smart phones." *37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN'07)*. IEEE, 2007.
3. Kouris, Ioannis, et al. "Mobile phone technologies and advanced data analysis towards the enhancement of diabetes self-management." *International Journal of Electronic Healthcare* 5.4 (2010): 386-402.
4. Ghahramani, Mohammadhossein, MengChu Zhou, and Chi Tin Hon. "Mobile phone data analysis: A spatial exploration toward hotspot detection." *IEEE Transactions on Automation Science and Engineering* 16.1 (2018): 351-362.
5. Elias, Daniel, et al. "SOMOBIL–improving public transport planning through mobile phone data analysis." *Transportation Research Procedia* 14 (2016): 4478-4485.
6. Ghahramani, Mohammadhossein, MengChu Zhou, and Gang Wang. "Urban sensing based on mobile phone data: approaches, applications, and challenges." *IEEE/CAA Journal of Automatica Sinica* 7.3 (2020): 627-637.
7. Bajardi, Paolo, et al. "Unveiling patterns of international communities in a global city using mobile phone data." *EPJ Data Science* 4 (2015): 1-17.
8. Freeland, S. L., and B. N. Handy. "Data analysis with the SolarSoft system." *Solar Physics* 182.2 (1998): 497-500.
9. Tukey, John W. "The future of data analysis." *The annals of mathematical statistics* 33.1 (1962): 1-67.
10. Ott, R. Lyman, and Micheal T. Longnecker. *An introduction to statistical methods and data analysis*. Cengage Learning, 2015.
11. Kruschke, John K. "What to believe: Bayesian methods for data analysis." *Trends in cognitive sciences* 14.7 (2010): 293-300.
12. Sheiner, Lewis B. "The population approach to pharmacokinetic data analysis: rationale and standard data analysis methods." *Drug metabolism reviews* 15.1-2 (1984): 153-171.
13. Scott, S. J., R. A. Jones, and WAI Williams. "Review of data analysis methods for seed germination 1." *Crop science* 24.6 (1984): 1192-1199.
14. Bryk, Anthony S., and Stephen W. Raudenbush. *Hierarchical linear models: Applications and data analysis methods*. Sage Publications, Inc, 1992.
15. Van de Vijver, Fons JR, and Kwok Leung. *Methods and data analysis for cross-cultural research*. Vol. 116. Cambridge University Press, 2021.
16. Weerahandi, Samaradasa. *Exact statistical methods for data analysis*. Springer Science & Business Media, 2003.
17. Höppner, Frank, et al. *Fuzzy cluster analysis: methods for classification, data analysis and image recognition*. John Wiley & Sons, 1999..
18. Bradburn, Mike J., et al. "Survival analysis part II: multivariate data analysis–an introduction to concepts and methods." *British journal of cancer* 89.3 (2003): 431-436.
19. Powers, Daniel, and Yu Xie. *Statistical methods for categorical data analysis*. Emerald Group

Publishing, 2008.

20. Gnanadesikan, Ram. *Methods for statistical data analysis of multivariate observations*. John Wiley & Sons, 2011.

21. Clifford, Gari D., and Francisco Azuaje. *Advanced methods and tools for ECG data analysis*. Ed. Patrick McSharry. Vol. 10. Boston: Artech house, 2006.