Smartphone price prediction

The continued expansion of the mobile phone market is inevitable. In the case of mixed mobile phone parameters, this project hopes to predict the price of mobile phones by establishing a model between mobile phone prices and various parameters. The purpose of this project is to determine the model that is easiest to predict the price of a mobile phone, and the parameters that have the greatest impact on the price of a mobile phone under this model.

1. Data

The dataset used in this project are all provided by Kaggle. This data is about the selling price and parameter information of Android phone. The main parameters are memory, running speed, mobile phone screen size and resolution, weight and so on. In this project, the price of the mobile phone will be the dependent variable, and other parameters will be the independent variable. The dataset is on the link below:

Mobile Price Prediction | Kaggle

2. Data cleaning and Pre-processing

The dataset provides 20 features of smartphone. The currency of target variable is the Indian rupee. The dataset contains currency mark, operate symbols, punctuation, and missing values. The goal of data cleaning is fill in all missing values, remove unnecessary symbols, and classify the data into numerical and category variables for subsequent predictive models.

- **Step 1**: cleaning all unnecessary symbols

  After checking the missing values in the dataset, I find many values are with symbols, such as price with "₹", resolution with unit "Pixels" and punctuation "x". They are not needed in arithmetic operations.
  Solution: use replace, apply method separating and removing symbols from value
- **Step 2**: Recreate the value containing the operation into a new column

Resolution is composed of height resolution and width resolution, and the value of display size contains both inches and centimeters.

Solution: create each value into a new column, and calculate their product respectively to form a new resolution column and display size column

- **Step 3**: separate numerical variables and categorical variables
  To define which variables will be used in the prediction model, I create two new data frames contain numerical and categorical variable respectively. The problem I find here is many of categorical variable are consist by Multiple strs with many different final values, even if they are all recombined, will still retain more than 10 different categorical type values. If all categorical variables are added to the prediction model, it will cause a huge increase in computational complexity and overfitting. Therefore, I decide that only the values of the numerical data frame are kept for subsequent prediction models.

The data cleaning and pre-processing is on the link below:

Data cleaning report

Data pre-processing report

3. Training the data with different model

The model used in the project listed below:

Linear Regression	k Neighbors Regression
Lasso Regression	SVM (support vector machine)

Ridge Regression Gradient Boosting

Random Forest Regression

I am going to find which model is the best model for predict smartphone price in there 7 model. I get the accuracy score, mean squared error and root mean square error to see which one gets the highest performance. Then I use grid search method to apply hyperparameter to improve the accuracy.

4. Model result summary

The graph below shows the result of each model's performance. The Random Forest Regression model has 93% accuracy rate and very low RMSE. The accuracy is much higher than other model, hence, we will take Random Forest Regression model as the main predict model. At next step, we will use grid search and assign hyperparameter to modify the model performance again.

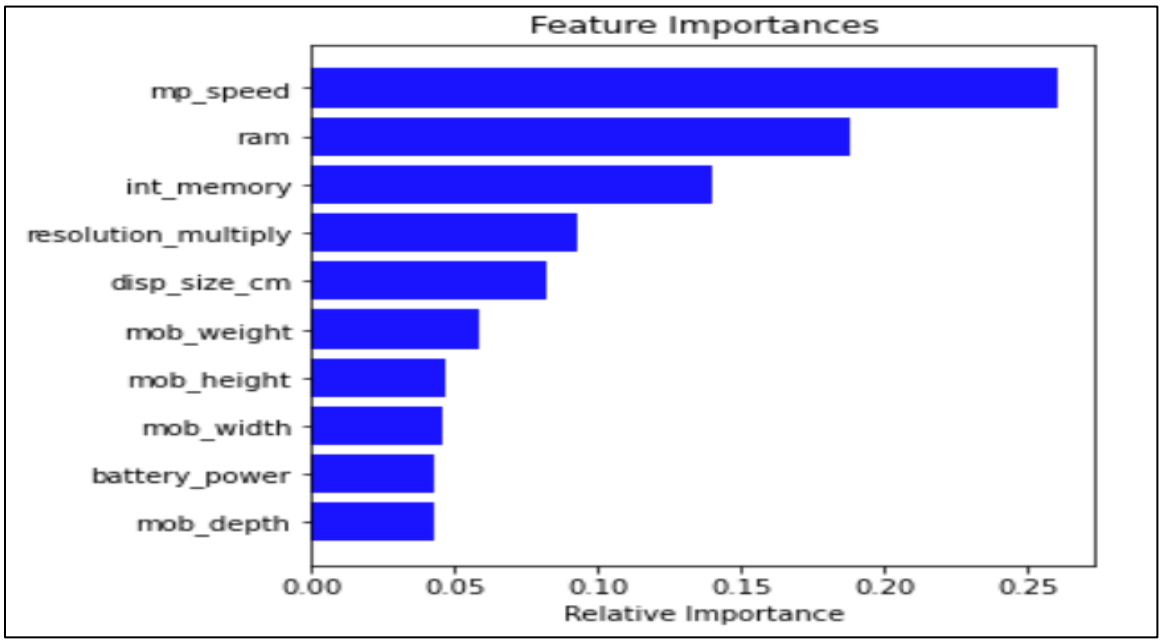
Model	Accuarcy score	MSE	RMSE
LogisticRegression	0.697938	2.20492e+07	4695.66
LassoRegression	0.697977	2.20463e+07	4695.35
RidgeRegression	0.695263	2.22444e+07	4716.4
RandomForestRegressor	0.946954	3.87212e+06	1967.77
KNeighborsRegression	0.757965	1.76675e+07	4203.27
SVM	-0.0876213	7.93916e+07	8910.19
Gradient Boosting	0.868836	9.57437e+06	3094.25

The graph below shows the score of random forest regression model after use hyperparameter. The accuracy score is 0.950477, RMSE is 1967.77. This model has 95% accuracy rate for predict smartphone price by assign parameters.

Model	Accuarcy score	MSE	RMSE
RandomForestRegressor	0.946954	3.87212e+06	1967.77
RandomForestRegressor_gird	0.950477	3.87212e+06	1967.77

5. Features importance

After I get the best prediction model get the feature importance level. The result is showed as below:



The results clearly show that the most important factor in determining the price of a phone is the performance of its processor, from speed to cache. They involve the core functions of the phone, and the Central Processing

Unit is still the most important than other parameters that affect appearance (resolution, height, weight) or battery life (battery power).

6. Future Improvements

- The operating system of all mobile phones in this project is Android. For iPhone, the world's largest phone brand, it could have turned out differently.
- In the future, comparing the importance of the same parameter under different operating systems may be a topic.

7. Credits

Thanks to my mentor Yadunath Gupta give me excellent advise and his mature data understanding.

Thanks to Dipanjan Sarkar and Kenneth Gil-Pasquel from springboard mentor.