Cell Phone Price Prediction [Draft White Paper]

- **Business Problem**

Cell Phone are one of the most widely used gadgets in recent years, every year there are some new feature upgrades happening and price increases accordingly. I am building a predictive model, that can predict the price category of cell phones depending on its features. In this problem the phone prices are divided into 4 categories ‘low cost’, ‘medium cost’, ‘high cost’ and ‘very high cost’. If the model works perfectly with good accuracy, it can be used for cell phone price predictions.

- **Background/History**

I have found the data source from Kaggle, the data has several features, which seems quite helpful and could be a good estimator for mobile price.

- **Data Explanation (Data Prep/Data Dictionary/etc)**

The data has two parts of it, one is training and other is test, there are several attributes in the data, I have declared them in the below list.

battery_power : Total energy of the battery can be stored(mAh)
blue : the device has blutooth or not, yes/no (1/0)
clock_speed : the speed of execution of instructions by microprocessor
dual_sim : the device has two SIM card at the same time or not
fc : the quality of front camera in MegaPixel
four_g : the device has 4G network or not
int_memory : internal memory in GigaByte
m_dep : the device depth in CM
mobile_wt : the weight of device
n_cores : the number of processor cores
pc : the quality of primary camera in MegePixel

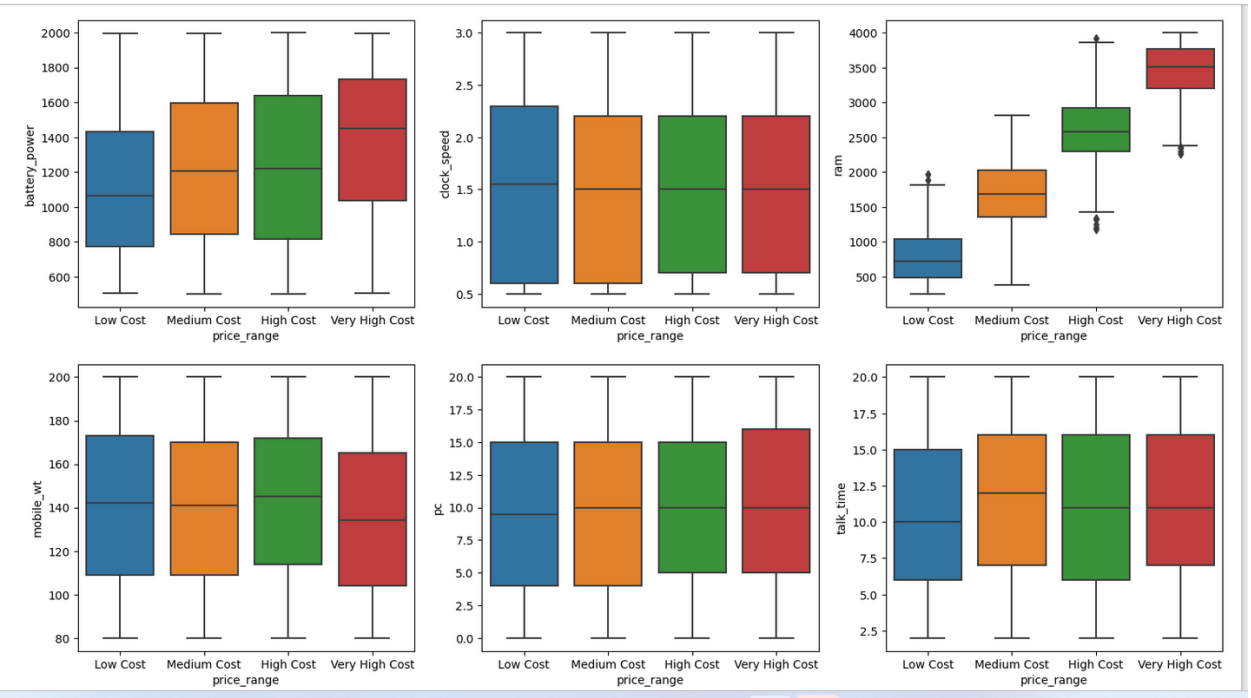
px_height : the height of pixel resolution
px_width : the width of pixel resolution
ram : random access memory in MegaByte
sc_h : the height of device screen in CM
sc_w : the width of device screen in CM
talk_time : the maximum talk time that the full charge battery of the device can support
three_g : the device has 3G network or not
touch_screen : the device has touch screen or not
wifi : the device has wifi or not
price_range : the categorized price of the device

- **Methods**

Here in this problem, I am trying to detect the mobile price category, as per the data, there are 4 categories of mobile prices, so it is a multiclass classification problem. I tried to complete different visual analysis to find out what features of mobile phone have influence on the cell phone price. I have also tried to fit two models to compare their accuracy of fit.

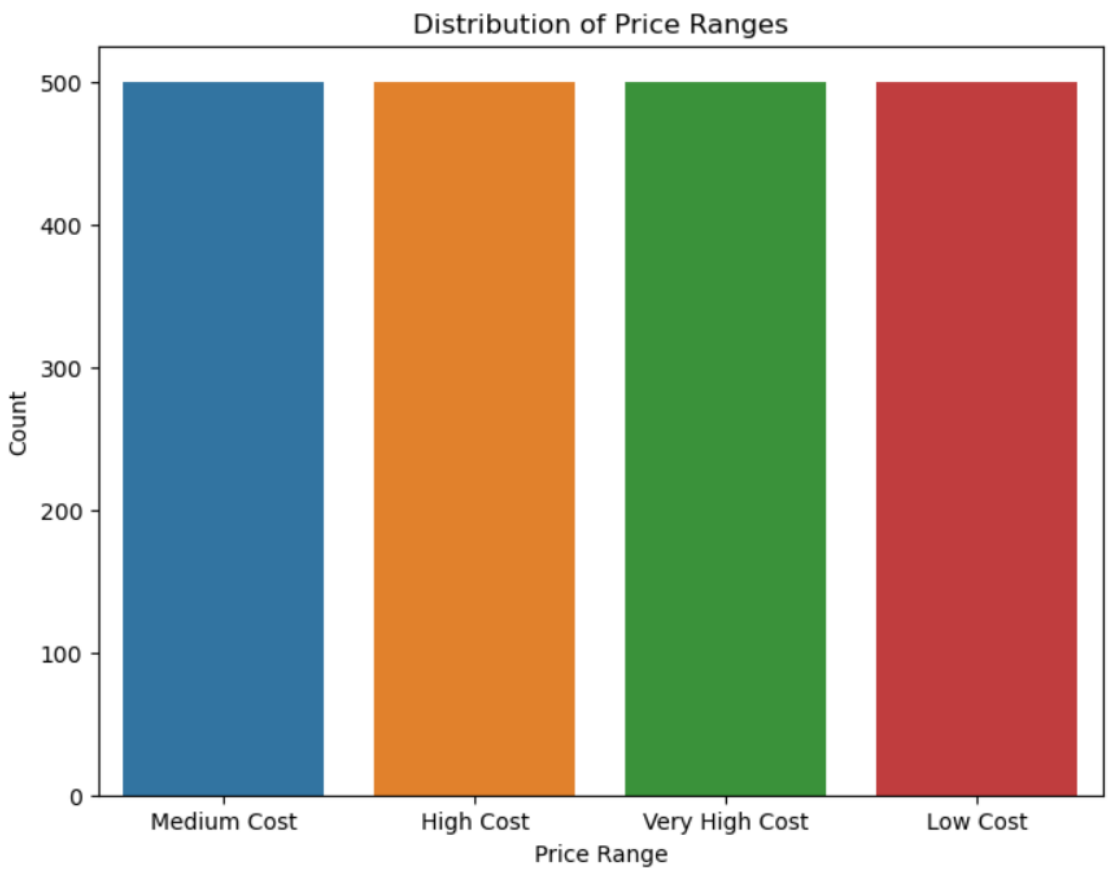
- **Analysis**

Initially I tried to find the count of missing records and found there is no missing records, so no imputation is needed. I did some bar plot on the continuous field to determine if those have influence on cell phone price.

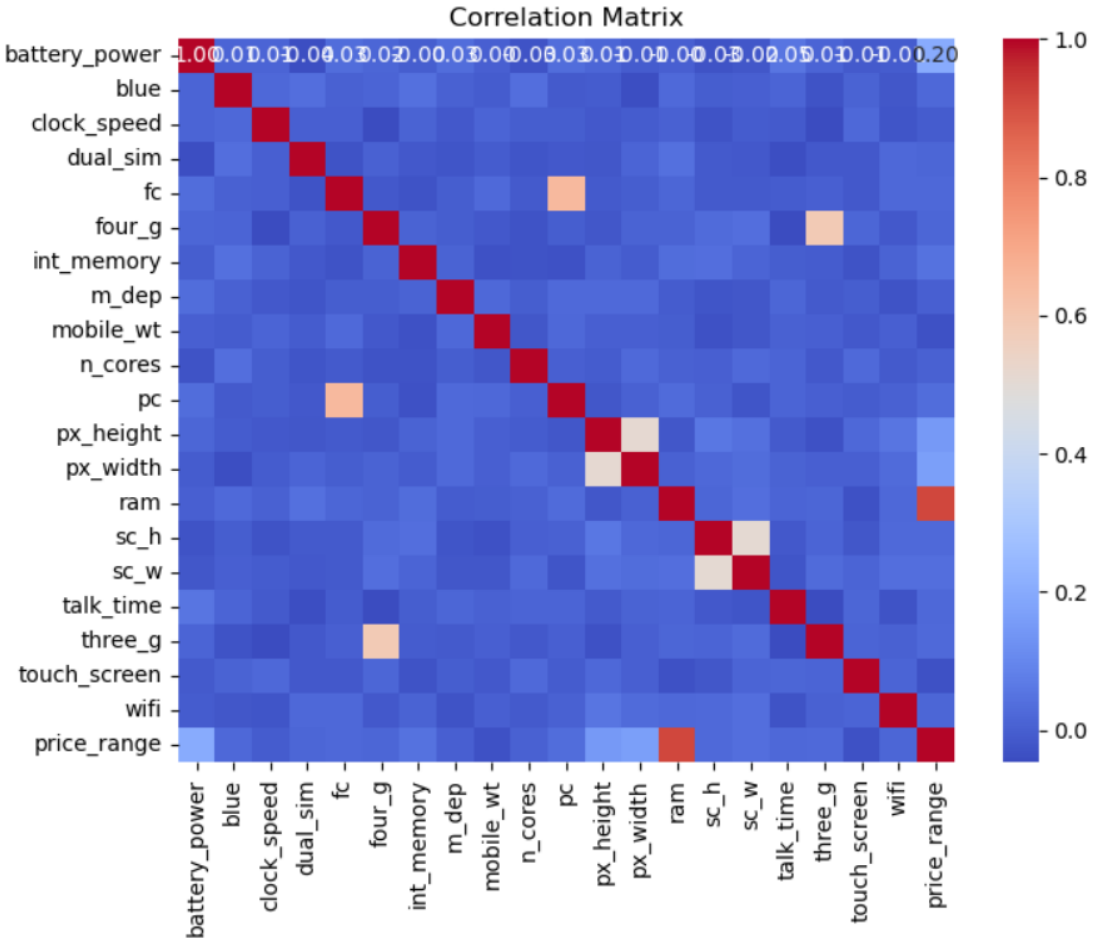


Looking at the above bar plots it clearly says the two features battery power and ram have major influence on cell phone prices. The higher battery power cell phones are more costly compared to lower battery power. Similar pattern is there for ram, the speed of the mobile is very much dependent on ram, higher ram has high cell phone prices and mobile having lower size rams are low in price.

There are four price categories and data has been created by taking uniform samples from each category, so we can say the prices in the sample of cell phone data are uniformly distributed as shown in the below diagram.



Below is the correlation matrix diagram on the cell phone data. The diagram also shows that ram has a high correlation with cell phone prices, i.e. the higher the ram the prices are higher for cell phones.



- **Conclusion**

I have used Random Forest Classifier and Decision Tree Classifier to predict the cell phone prices, i.e. in which category out of four the cell phone prices belon to. I have used the 20% test sample to test the accuracy of the models, out of the two fitted models it came out that the Random Forest Classifier is more accurate compared to the Decision Tree Classifier.

- **Assumptions**

The only assumption is that the encoded integer value for each variable should have an ordinal relation.

- **Limitations**

The major limitation I would say for this prediction model is data, there are several other key features on the camera, that is not used in this model prediction. In general people use

cameras from cell phones and several camera features have been introduced in these recent days, that is lagging in this data set. I would say there should be more camera features that need to be used in model fitting to predict the prices more accurately. Another major limitation is price category, the cell phone prices vary in wide range, in this dataset the price categories are low, medium, high and very high from the categories it does not give clear idea about the prices for different cell phones.

- **Challenges**

There are several challenges for the application that can predict the cell phone price. Due to technological change, every week there are some new features being introduced or existing features have a major change. There should be re-calibrated very frequently for better prediction

- **Future Uses/Additional Applications**

This phone price prediction model can be upgraded in future by adding several front and rear camera related features, which is one of the major attractions in today's cell phone. We can update the model and push the updated model pickle files to the application for better predication.

- **Recommendations**

I have fitted two models in the data to compare the accuracy, one is Random Forest Classification model, and the other is Decision tree Classification model.

Accuracy for Random Forest Model

[112]:
```
# Make predictions on the test set
y_pred = rf_classifier.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)

# Generate classification report
print(classification_report(y_test, y_pred))
```

```
Accuracy: 0.8925
              precision    recall  f1-score   support

           0       0.95      0.96      0.96       105
           1       0.89      0.87      0.88        91
           2       0.78      0.87      0.82        92
           3       0.94      0.87      0.90       112

    accuracy                           0.89       400
   macro avg       0.89      0.89      0.89       400
weighted avg       0.90      0.89      0.89       400
```

Accuracy for Decision Tree Model

[114]:
```
# Make predictions on the test set
y_pred = dt_classifier.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)

# Generate classification report
print(classification_report(y_test, y_pred))
```

```
Accuracy: 0.835
              precision    recall  f1-score   support

           0       0.93      0.86      0.89       105
           1       0.74      0.86      0.79        91
           2       0.79      0.72      0.75        92
           3       0.88      0.89      0.89       112

    accuracy                           0.83       400
   macro avg       0.83      0.83      0.83       400
weighted avg       0.84      0.83      0.84       400
```

Among the two models it came out that Random Forest model is more accurate in this case compared to the Decision Tree model.

- **Implementation Plan**

I am planning to build a web application that can be used to re-sell used cell phones. Recently people did several phone upgrades and are reselling their old phones. We can build an API that can be called from different platforms and other applications with features of cell phones and the API will return the cell phone price category to suggest cell phone prices to the customer. Those prices will help customers to list their used phones.

- **Ethical Assessment**

While implementation the prediction as an API, we must be careful on several thing, as user will input about the mobile information and different feature related information will be available on the server for all the user's personal mobile. So, the data must be purged in a timely manner. We must be careful not to collect user personal information tagged with the mobile as the scope of the application is only limited to cell phone price prediction.