Business Presentation Project Recell

By Minh Ngo

Table of contents

Content
Business Problem Overview and Solution Approach
Data Preprocessing
Data analysis - EDA
Univariate analysis
Multivariate analysis
Model summary
Key insights & Conclusion

Business Problem Overview and Solution Approach

- **Core business idea**
  - Used and refurbished smartphones is growing very fast due to various factors (Covid19, customers cutting back on spending, longer longevity of smartphone etc.)
  - Recell wants to tap into this market and grow as a professional phone reseller
  - Recell needs to understand the impact of different factors on the final used prices so that they can have a better pricing strategy
- **Problem to tackle**
  - Recell wants to use ML-based solution to **develop a dynamic pricing strategy** for used and refurbished smartphones.
- **Financial implications**
  - Based on the market data, Recell would understand the prices that customers are willing to pay, therefore they can set the prices to **maximize revenue**
  - They can also **maximize profit** by developing a threshold for the prices of used phones that they will buy , so that they won’t over pay for any phone they buy, Therefore they can **minimize the cost of good sold** and maximize profit

Business Problem Overview and Solution Approach

- How ML model can solve the problem
  - In this project we will use Linear Regression model to predict the phone prices. However this is not a simple linear regression. We will apply Machine learning , particularly Supervised Learning in the process
  - With Regression – Supervised learning, we would split the data set into 2 sub dataset and learns from the training data using these target variable as reference variable. The model generated would then be used to make predictions about the data to see the model before
  - With this method, the model **can learn from the data and generate a line that best fits the data**.
  - This line is basically the regression – the model that we use to **predict the prices** from all available variables

Data Overview

- Brief description of data provided

Observation	Variable
3571	15

Note:
- There are some missing values from the dataset, we will review carefully later during the data processing part
- There are several object datatype (brand_name, os, 4g,5) which were converted into categorical for better preprocessing

Variable	Number of missing value
main_camera_mp	180
selfie_camera_mp	2
int_memory	10
ram	10
battery	6
weight	7

Data Overview

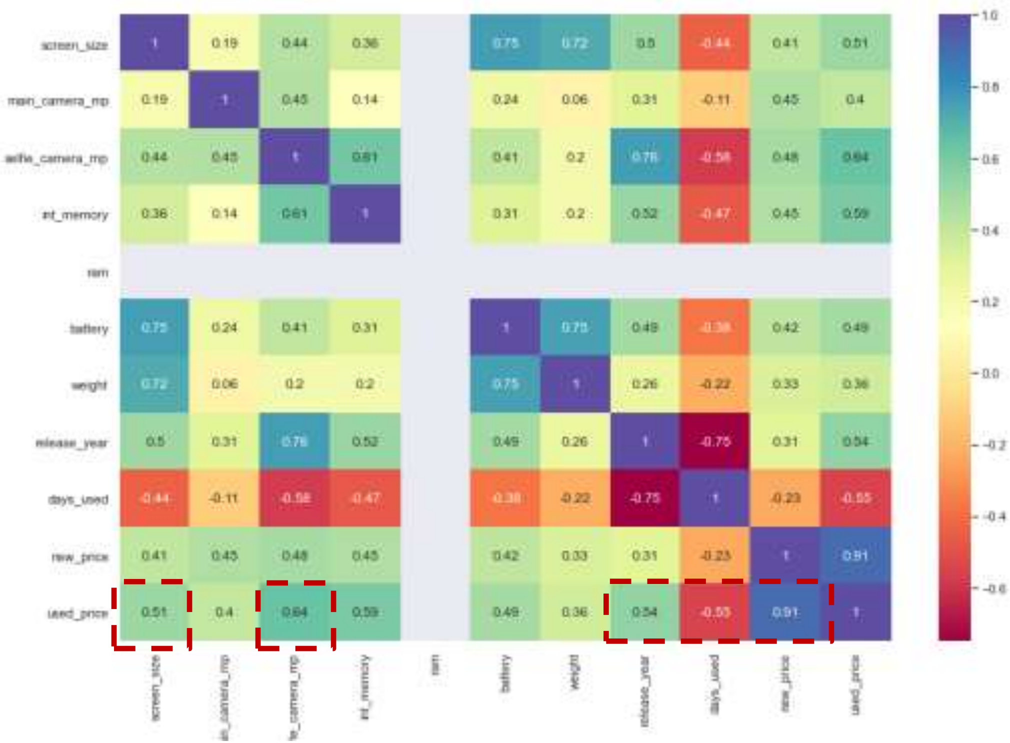
- Brief description of data provided

#	Variable	Description
1	brand_name	The brand name of each phone
2	os	Type of operating system
3	Screen_size	Size of screen in cm
4	4G	Whether 4G is available or not
5	5G	Whether 5G is available or not
6	Main camera mp	Resolution of main camera in pixel
7	Selfie camera mp	Resolution of main camera in pixel
8	Int memory	Internal memory of the phone
9	Ram	Amount of ram in GB
10	Battery	Energy capacity of the phone battery in mAh
11	Weight	Weight in gram
12	Release_year	Year when the phone model was released
13	Days_used	Number of days that the used phone has been used
14	New_price	Price of a new phone of the same model in euros
15	used_price	Price of a used phone in euros

Data preprocessing

- Below are significant data preprocessing steps that were made to the raw data
  - **Missing value treatment:** There are some value missing from different columns, we replaced the missing value with the median value
  - **Outlier treatment:** there are high outliers in screensize, main camera mp, selife camera mp, int memory, batter, weight, new price and used price. We treated these outliers by **flooring** and **capping**. After treatment, there are no longer any outliers in the dataset.

EDA
Correlation heat map



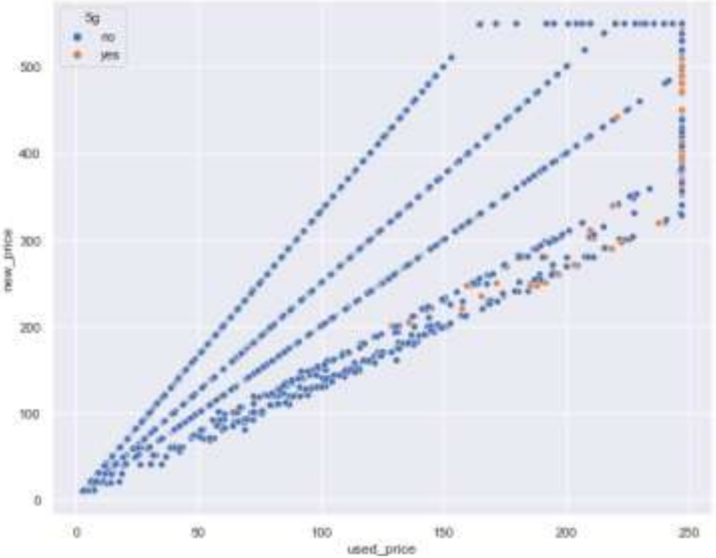
Observation

Key factors that have high impact on used prices are: **selfie camera, screen size , release year, days used, new price**

- Used price and **New price** are highly correlated (correlation = 0.93, indicating that the used price is strongly dependent on the new price)
- Used price is **negatively correlated** with **days used** (meaning the more days the phone is used, the lower the price - which also makes sense)
- Used price also **positively correlated** with **selfie camera mp** (correlation =0.5) and **ram** (correlation = 0.52)
- **Released year** has correlation of 0.46 with used price

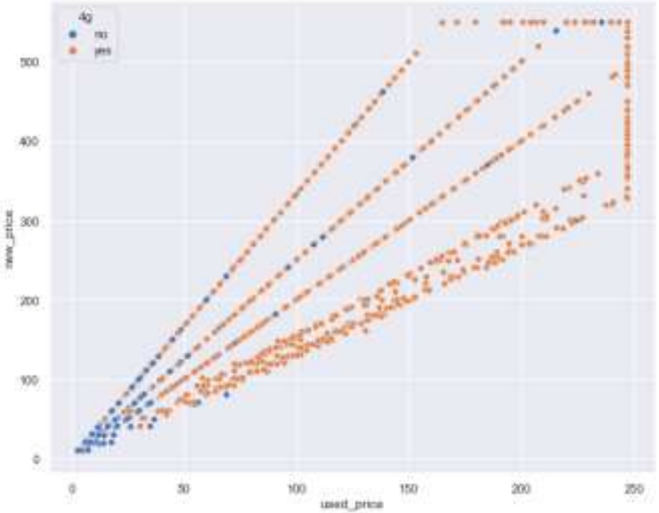
EDA
Bivariate analysis

Used price vs new price vs 5G status



Used price has very strong correlation with new price, yet the correlation is even stronger for phone with 5G.
This means that for any phone having 5G, its new price has a stronger predictive power of used price

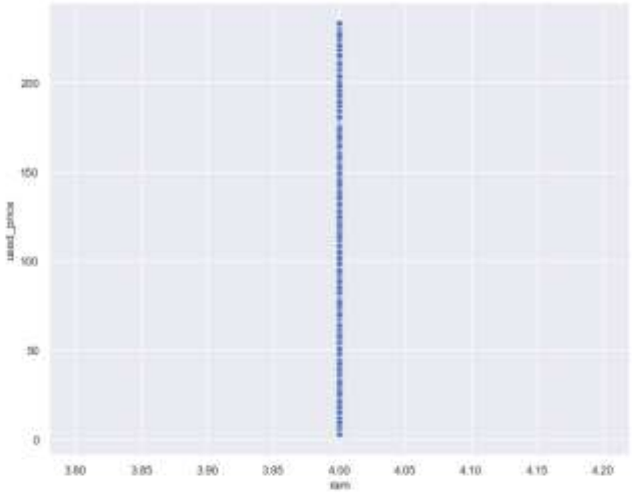
Used price vs new price vs 4G status



The correlation between used price and new price is stronger for phones with 4G (including 5G) status.

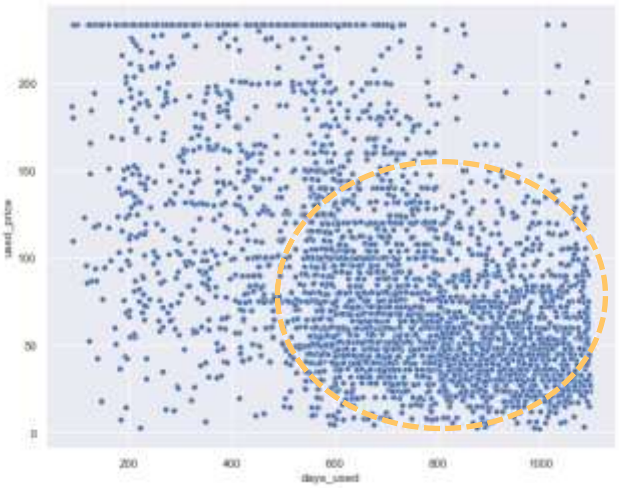
EDA
Bivariate analysis

Ram vs used_price



When outliers were removed, only ram of 4mb are left in the dataset

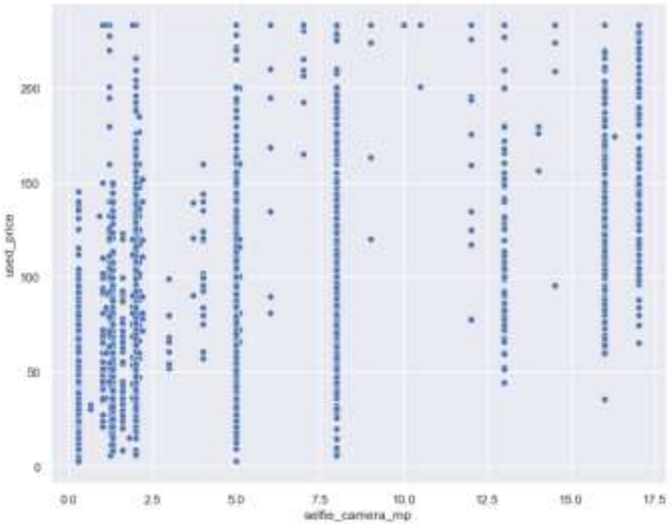
Days_used vs used_price



Phone with lower number of days_used (mostly from 200-450 days) are more likely to have higher used price

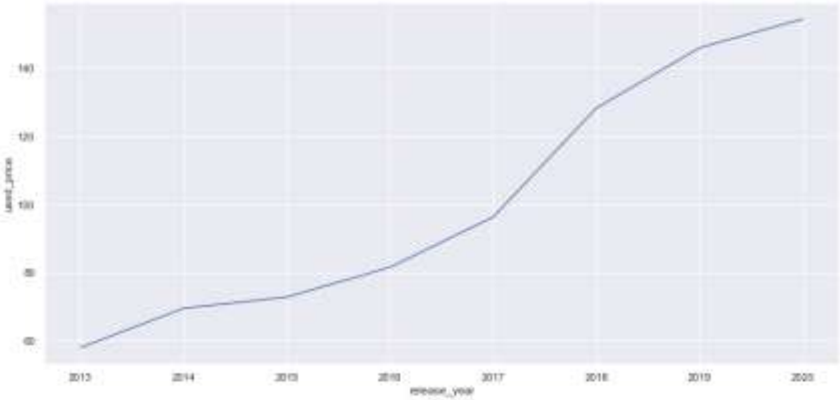
EDA
Bivariate analysis

Selfie_camera_mp vs used_price



Phones that have selfie_camera_mp from 15-25 are likely to have higher prices

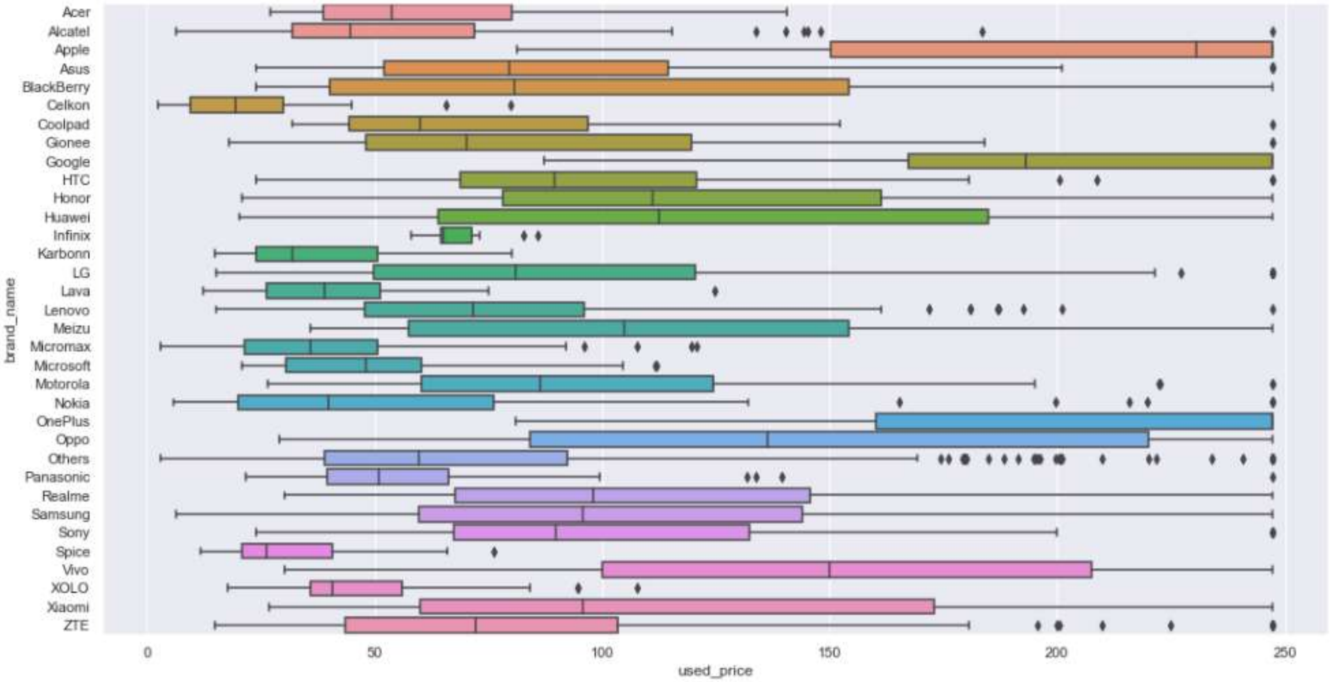
Released_year vs used_price



* phones released from 2017 onward see the significant increase in price compared with phones released before that

EDA
Bivariate analysis

Brandname vs used_price



From this chart we can see the range of used phone of different brand

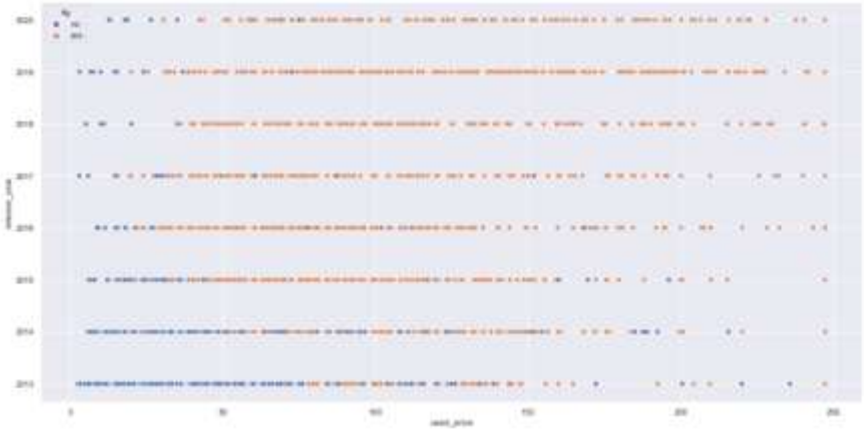
Acer has smallest price range (40-150USD)

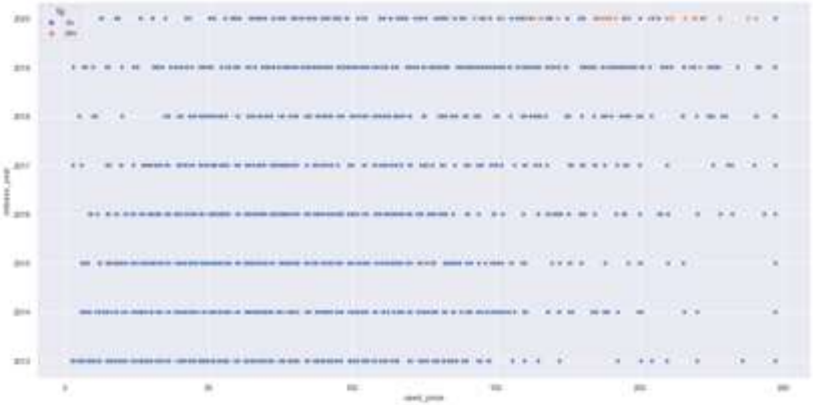
Most phones are sold within price range of 50-100

Apple, Google and Oneplus have highest price distribution

EDA
Bivariate analysis

4G vs release year vs used_price





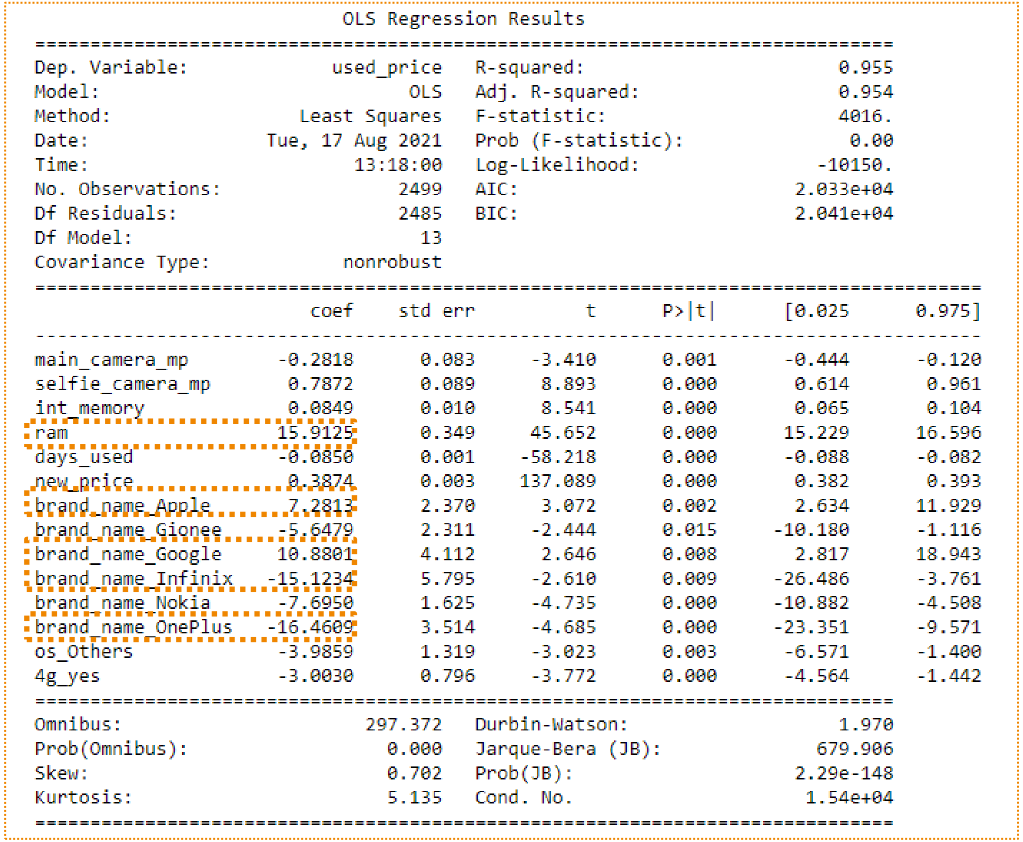
* Most of the phones without 5G were released during period of 2013-2016 and they have lower prices (from 10-50USD) than phone with 5G

Most of the phones with 5G were released in 2020 and they have higher prices than phone without 5G

Model Performance Summary

- Overview of ML model and its parameters
  - The ML model is supervised learning – linear regression and we look for the linear relationship between all other factors vs used price
  - After we refined the model, All of the assumptions of linear regression were met

- The model's parameters:

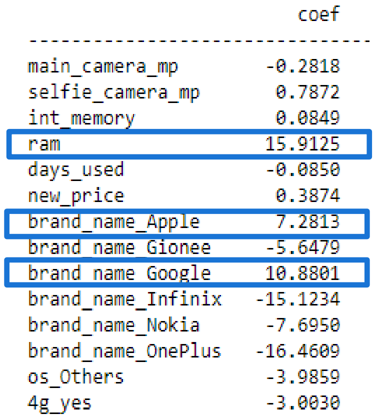


```
                            OLS Regression Results
==============================================================================
Dep. Variable:             used_price   R-squared:                       0.955
Model:                            OLS   Adj. R-squared:                  0.954
Method:                 Least Squares   F-statistic:                     4016.
Date:                Tue, 17 Aug 2021   Prob (F-statistic):               0.00
Time:                        13:18:00   Log-Likelihood:                -10150.
No. Observations:                2499   AIC:                         2.033e+04
Df Residuals:                    2485   BIC:                         2.041e+04
Df Model:                          13
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
main_camera_mp      -0.2818      0.083     -3.410      0.001      -0.444      -0.120
selfie_camera_mp     0.7872      0.089      8.893      0.000       0.614       0.961
int_memory           0.0849      0.010      8.541      0.000       0.065       0.104
ram                 15.9125      0.349     45.652      0.000      15.229      16.596
days_used           -0.0850      0.001    -58.218      0.000      -0.088      -0.082
new_price            0.3874      0.003    137.089      0.000       0.382       0.393
brand_name_Apple     7.2813      2.370      3.072      0.002       2.634      11.929
brand_name_Gionee   -5.6479      2.311     -2.444      0.015     -10.180      -1.116
brand_name_Google   10.8801      4.112      2.646      0.008       2.817      18.943
brand_name_Infinix -15.1234      5.795     -2.610      0.009     -26.486      -3.761
brand_name_Nokia    -7.6950      1.625     -4.735      0.000     -10.882      -4.508
brand_name_OnePlus -16.4609      3.514     -4.685      0.000     -23.351      -9.571
os_Others           -3.9859      1.319     -3.023      0.003      -6.571      -1.400
4g_yes              -3.0030      0.796     -3.772      0.000      -4.564      -1.442
==============================================================================
Omnibus:                      297.372   Durbin-Watson:                   1.970
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              679.906
Skew:                           0.702   Prob(JB):                    2.29e-148
Kurtosis:                       5.135   Cond. No.                     1.54e+04
==============================================================================
```

Model Performance Summary

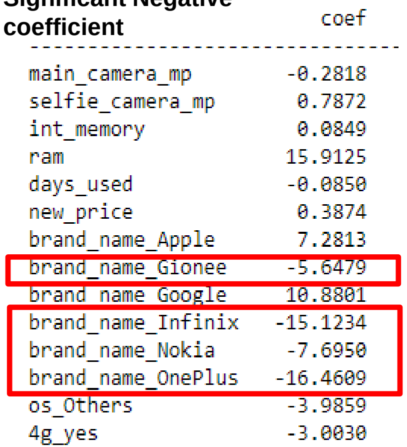
- The most important factors used by the ML for prediction is the coefficient : coefficient tell us how much the used price (dependent variable) is expected to increase (if coef is positive) or decrease (if coef is negative) when that independent variable increase by one. Below are the most important factors from the ML model that can predict the used price

Significant Positive coefficient



```
                               coef
------------------------------
main_camera_mp             -0.2818
selfie_camera_mp            0.7872
int_memory                  0.0849
ram                        15.9125
days_used                  -0.0850
new_price                   0.3874
brand_name_Apple            7.2813
brand_name_Gionee          -5.6479
brand_name_Google          10.8801
brand_name_Infinix        -15.1234
brand_name_Nokia           -7.6950
brand_name_OnePlus        -16.4609
os_Others                  -3.9859
4g_yes                     -3.0030
```

Significant Negative coefficient



```
                               coef
------------------------------
main_camera_mp             -0.2818
selfie_camera_mp            0.7872
int_memory                  0.0849
ram                        15.9125
days_used                  -0.0850
new_price                   0.3874
brand_name_Apple            7.2813
brand_name_Gionee          -5.6479
brand_name_Google          10.8801
brand_name_Infinix        -15.1234
brand_name_Nokia           -7.6950
brand_name_OnePlus        -16.4609
os_Others                  -3.9859
4g_yes                     -3.0030
```

•Holding all other features fixed, one unit increase in **Ram** is associated with an **increase of $15.9 in used price**

•Holding all other features fixed, one unit increase in **Apple brand** is associated with an **increase of $7.2 in used price**

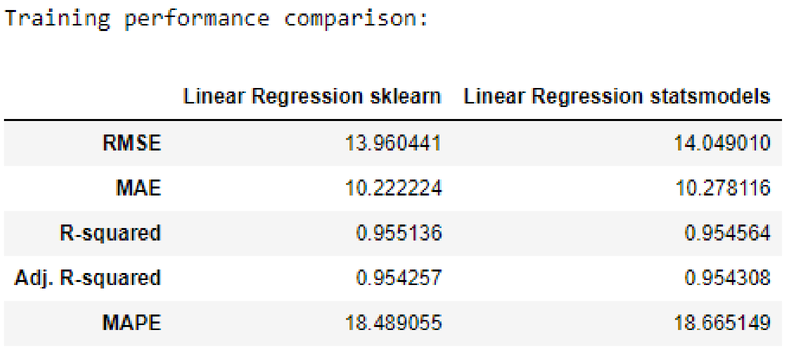
•Holding all other features fixed, one unit increase in **Google** is associated with an **increase of $10.8 in used price**

•Holding all other features fixed, one unit increase in **OnePlus** is associated with a decrease **of $16.4 in used price**

•Holding all other features fixed, one unit increase in **Infinix** is associated with a decrease **of $15 in used price**

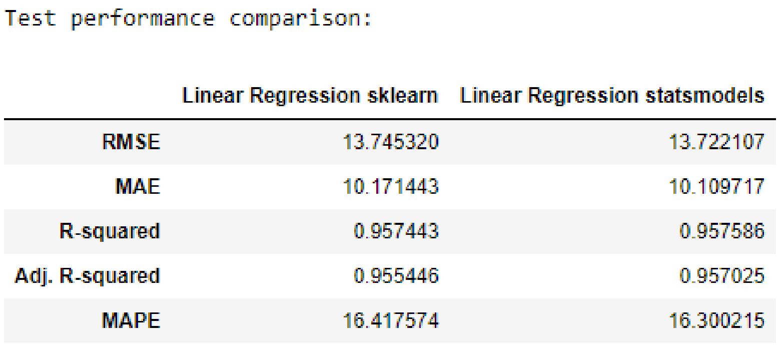
Model Performance Summary

- Below are key performance metrics for training and test data. We can see that the performance of two models are close to each other



```
Training performance comparison:
```

	Linear Regression sklearn	Linear Regression statsmodels
RMSE	13.960441	14.049010
MAE	10.222224	10.278116
R-squared	0.955136	0.954564
Adj. R-squared	0.954257	0.954308
MAPE	18.489055	18.665149



```
Test performance comparison:
```

	Linear Regression sklearn	Linear Regression statsmodels
RMSE	13.745320	13.722107
MAE	10.171443	10.109717
R-squared	0.957443	0.957586
Adj. R-squared	0.955446	0.957025
MAPE	16.417574	16.300215

Business Insights and Recommendations

Insights

- All factors that have positive impact on price are: Ram, Google, Apple, selfie camera, new price and internal memory. As these factors increase, the used price increase

- Ram and Google and Apple brand name turns out to have a **very significant impact** on the price of used phones. As these factors increase, the use price increase (as these two have positive relationship with used phone price, and they have positive coefficient sign)

- Of all brands, here are the brands that have most positive impact on the used prices: **Apple, Google**
- Of all brands, here are the brands that have negative impact on the used prices: Nokia, Infinix Gionee, OnePlus.
- Infinix and OnePlus have a strongest negative impact on the used price with coefficient being -15 and -16 . This mean if the phone is under these 2 brands, the price will decrease by 15 or 16 USD consecutively

Recommendation

- Based on the model, Recell would know how to price different phones, using different factors and attributes of the phone.
- Recell would want to be **selective** when **stocking** their inventory or pricing the product
- If they want to increase revenue by selling phones with higher prices, they can **stock top brands** that can impact the prices positively, namely Google, Apple. Also they want to carry phones with large RAM and selfie camera resolution
- They **should not carry** a lot of phones under Oneplus, Nokia, Gionee, Inifinix. As these brands have negative impact on prices.
- Phones with larger ram capacity would also indicate higher prices so they **should charge higher for phones with larger RAM**

Business Insights and Recommendations

- **Data source for model improvement** :
  - **Potential datapoint and features to be included in the model**: Product features (such as waterproof, shock proof), other variables such as Dimension, CPU, warranty period (one important indicator of phone quality)
  - **Some features that could be improved:** We should consider not to split 4G and 5G into 2 columns because 4G already include 5G, instead we should have 1 column that have 4G, 5G and others. Currently it is misleading to have 4G include 5G so 5G is double counted in 4G column,
- **Model implementation in real world:** Since the R-square and adjusted R-squared are very high (95%) for both test and train dataset, we can be confident that this model works well and can be deployed within ReCell.

Business Insights and Recommendations

- **Potential business benefits from model:**
  - By applying the model in pricing strategy, ReCell can have an **upper hand** in **positioning** / pricing the products in reference to other options on the market, based on the market data. A good pricing strategy would make the products more appealing to customers while covering the cost.
  - Recell would be able to **maximize revenue** (carrying higher price used phone) and **maximize profit** by reducing the cost (by not overpaying)
  - Linear regresion is a fairly **simple** model and can **be reused again** (using new data) to evaluate trends and make estimates of the price .

greatlearning

Power Ahead

Happy Learning !