ReCell

Supervised Learning, PGP - Data Science and Business Analytics

Arsalaan B. Saiyed

Date – 13th May 2022

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- Data Background and Contents
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendices
    - Appendix-A: EDA Results
    - Appendix-B: Model Performance Summary

Executive Summary

- Our study reveals that the used price of the phone is primarily dependent on the following:
  - Price of new device
  - Years since release (inversely proportional)
  - os - Operating system (inversely proportional)
  - ram - Processing speed
  - Main camera & front camera
  - Weight
- Based on the above findings from the given data we can provide the following insights and recommendations business up:
  1. The original price of the devices plays a huge role in the price of the used devices ReCell should aim to acquire as many devices that have a high market prices. For example, Apple, Oneplus and Google products are higher in quality and high in performance which can be bought at lower rates, refurbished and sold at comparatively higher prices than other brands.
  2. The older the product is the more outdated it gets and depreciates in value. Therefore, ReCell is much more likely to have success selling more recent phones/tablets at a competitive prices. Similarly, the lesser the device was previously used the better.
  3. Our model suggest that the ram and operating system of the device also contributes to the used price of the product. Any product which does not operate on android, IOS or windows is likely to have poor demand and resale value. Customers prefer faster devices so devices with lesser than 4gb ram are unlikely to be in demand.
  4. It is no surprise customers constantly want better quality camera. Better the camera quality the better the resale prices.

Executive Summary

5. ReCell can hold ad campaigns to target customer based on their needs and divide them into segments. And based on the demand in those segments they can stock up on the best devices at competitive prices.
6. Heavier phones tend to be priced higher as well probably because it gives a feeling of high quality. So devices with larger batteries would tend to be heavier hence, sell better.

- A include more columns few recommendations from outside the data:

1. We can in to our database as business grows over the years for better prediction. We can ask customers for ratings for things such as overall satisfaction of the used device, used price fairness: which rates how much do they feel the price for the used device is fair.
2. Market research on what makes people buy new devices because that drives people selling their old devices. Usually, people buy new devices for the following reasons:
    1. It was lost, turned faulty or was damaged – This is not a useful factor
    2. Devices with significantly better features came out – Tech enthusiasts always want the next best thing so we can capitalize with the devices that they let go off too early by buying from them and selling them to everyday people looking for a new device or a simple upgrade.
3. ReCell can partner up with big brands to get certified technicians that repair/refurbish old devices for ReCell and can be sold as a certified distributor.

Business Problem Overview and Solution Approach

- Problem Overview:

  The market for used and refurbished phones and tablets has been forecasted to grow with a compound annual growth rate (CAGR) of 13.6%. This growth can be attributed to an uptick in demand for used phones and tablets that offer considerable savings compared with new models.

  The objective is to provide an ML-based solution. To help ReCell tap in to the potential in this market we have built a linear regression model to predict the price of a used phone/tablet and to identify factors that significantly influence it.

- Solution Approach:

  - Data Overview & EDA
  - Data pre-processing
  - Linear Regression Model Building
  - Test of linearity & Final Model Summary
  - Executive Summary - Actionable insights and summary

Data Background and Contents

- Initial Observations:
  I. Data was formatted appropriately.
  II. Some of the data was missing for the columns highlighted in red, refer Table 2.
  III. All rows in the data were unique which means there was data for 3454 unique phones/tablets of 34 unique phone/tablet brands and 4 different operating systems.

Observations	Variables
3454	15

Table 1: Shape of the data

Variable	Description
brand_name	Name of manufacturing brand
os	OS on which the device runs
screen_size	Size of the screen in cm
4g	Whether 4G is available or not
5g	Whether 5G is available or not
main_camera_mp	Resolution of the rear camera in megapixels
selfie_camera_mp	Resolution of the front camera in megapixels
int_memory	Amount of internal memory (ROM) in GB
ram	Amount of RAM in GB
battery	Energy capacity of the device battery in mAh
weight	Weight of the device in grams
release_year	Year when the device model was released
days_used	Number of days the used/refurbished device has been used
normalized_new_price	Normalized price of a new device of the same model in euros
normalized_used_price	Normalized price of the used/refurbished device in euros

Table 2: Details

EDA Results – Univariate Analysis

Upon conducting a univariate analysis of the raw data and computing the statistical summary we observed the following:

I. Normalized new price is found to be at a mean of Euro 5.23

II. Normalized used price is found to be at a mean of Euro 4.36

III. Most of the devices run on an Android operating systems

IV. Most devices are 4g compared to 5g

V. Many older devices are not sold

VI. Most devices,502 to be exact are from brand name ‘others’ meaning the brand was unknown at the time of the data sourcing. The 2nd and 3rd most devices are from the brands Samsung and Huawei respectively.

Note: The mean, min, max etc. for all columns are given in the statistical summary in a table in Appendix-A

Link to Appendix slide on data background check

EDA Results – Bivariate Analysis

Upon conducting a bivariate analysis of the raw data we observed the following:

I. From the heatmap we can see that lot of device features have a positive correlation with the used prices.

II. Days used/release year i.e. the age of the devices have a negative correlation with the used prices.

III. We observe the wide range of variability of features across brand names.

IV. Some customers specifically look for good front cameras to click cool selfies. According our data Huawei, Vivo and Oppo offer most devices with front facing cameras with more than 8 megapixels

V. Some customers specifically look for good rear facing cameras to click cool selfies. According our data Sony and Motorola offer most devices with rear cameras with more than 16 megapixels.

VI. For customers that prefer high quality both rear and front cameras can go with Sony or HTC

VII. Customers that prefer to use devices mainly for entertainment or creative purposes prefer a larger screen size. They could go with brands such as Huawei and Samsung that provide the most devices with screens bigger than 6 inches.

VIII. Some customer need faster processing speed for this they can go for brands like Samsung, Oppo and Huwaei for rams greater than 8gb.

Link to Appendix slide on data background check

Data Preprocessing

- Duplicate value check
- Missing value treatment
- Outlier check (treatment if needed)
- Feature engineering
- Data preparation for modeling

Note: All pre-processing was carried out while conducting business sense checks along each step.

Data Preprocessing

- Let's take a quick look at a small sample of our data.
- This will help us get some idea of the attributes of the data and also help us understand the degree of cleaning needed before we can build a model.
- Duplicate Value Check: No duplicate entries were found in the given data

Random sample of the data:

	brand_name	os	screen_size	4g	5g	main_camera_mp	selfie_camera_mp	int_memory	ram	battery	weight	release_year	days_used	normalized_used_price	normalized_new_price
866	Others	Android	15.24	no	no	8.00	2.0	16.0	4.00	3000.0	206.0	2014	632	4.038832	5.190788
957	Celkon	Android	10.16	no	no	3.15	0.3	512.0	0.25	1400.0	140.0	2013	637	2.800325	3.884652
280	Infinix	Android	15.39	yes	no	NaN	8.0	32.0	2.00	5000.0	185.0	2020	329	4.370713	4.487287
2150	Oppo	Android	12.83	yes	no	13.00	16.0	64.0	4.00	3200.0	148.0	2017	648	4.677863	5.639422
93	LG	Android	15.29	yes	no	13.00	5.0	32.0	3.00	3500.0	179.0	2019	216	4.517650	5.300415
1040	Gionee	Android	12.83	yes	no	13.00	8.0	32.0	4.00	3150.0	166.0	2016	970	4.645640	5.634325
3170	ZTE	Others	10.16	no	no	3.15	5.0	16.0	4.00	1400.0	125.0	2014	1007	3.764451	4.244344
2742	Sony	Android	12.70	yes	no	20.70	2.0	16.0	4.00	3000.0	170.0	2013	1060	4.422809	5.799820
102	Meizu	Android	15.29	yes	no	NaN	20.0	128.0	6.00	3600.0	165.0	2019	332	4.959412	6.040659
1195	HTC	Android	10.29	no	no	8.00	2.0	32.0	4.00	2000.0	146.0	2015	892	4.227855	4.879007

Data Preprocessing

- Missing Value Treatment:

  Table below shows the number of missing values in all the columns in the given data. Upon analyzing them we conclude none of them can be dropped and are essential for the model building therefore, we will impute the data with their median.

Column	Missing Variable
main_camera_mp	179
weight	7
battery	6
int_memory	4
ram	4
selfie_camera_mp	2
brand_name	0
os	0
screen_size	0
4g	0
5g	0
release_year	0
days_used	0
normalized_used_price	0
normalized_new_price	0

I. We first impute the missing data with the median of the data grouped by brand name and release year to consider brand as well as temporal factor.

II. Interestingly we find that not all the missing values are imputed and this is because for some brands there is no data available for certain columns for certain years.

III. Therefore, we impute the missing data once more with the median of the data but grouped by brand name only and this results in the complete treatment of the missing values.

Data Preprocessing

- Outlier Check:

  On reviewing the boxplots we can see a lot of variability throughout many of the features but this is a result if the wide range of specifications brands try to offer from low end just for day-to-day or minimal use or high end products tailored for high performance for customers looking for a top of the line product. Some such occurrences can be seen in the following:

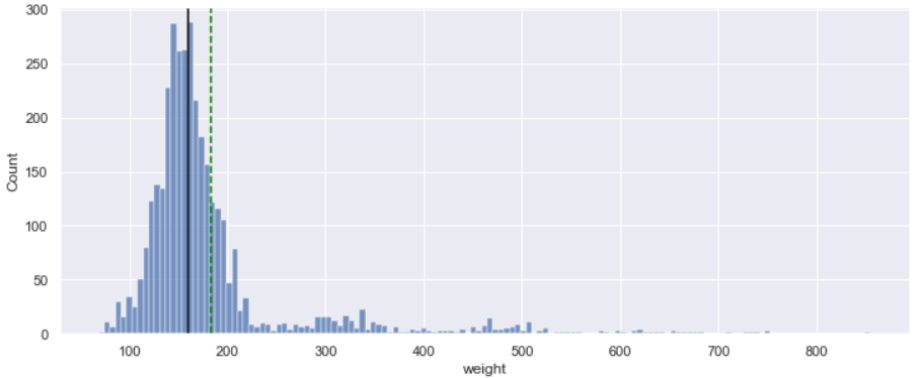
  - Front and rear cameras with over 20 megapixels. Phones such as Nokia, ZTE and Motorola do in fact provide these features.

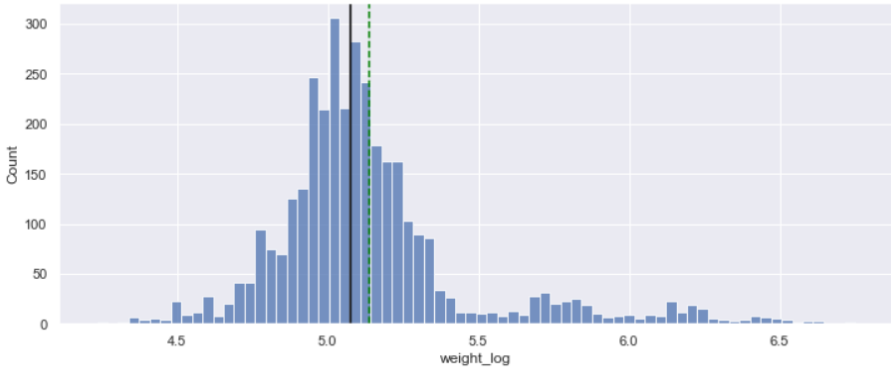
  - Internal memory over 500 GB is available in phones and tablets offered by brands such as Apple, Asus and Samsung

- In conclusion, no outlier treatment is required as these are not miss entries or impossible values. It is crucial for our model building to factor in all features available in the market.

Data Preprocessing

- Feature Engineering:
  - We create a new column years_since_release from the release_year column. We considered the year of data collection, 2021, as the baseline and then we dropped the release_year column.
  - The weight column is right skewed so to deal with this we using log transform to normalize the data. We create a new column called weight_log and drop the weight column. Refer graphs below
  - We create new columns for categorical columns like 4g, 5g, and os. We drop brand_name for our model as it doesn't add any value to model.





Model Performance Summary

- Overview of ML model and its parameters
- Summary of steps and most important factors used by the ML model for prediction
- Model Assumptions
- Summary of key performance metrics for training and test data in tabular format for comparison

***Note**: You can use more than one slide if needed*

Link to Appendix slide on model assumptions

Model Performance Summary

- Overview of ML model and its parameters:

  Our goal was to provide ReCell with a smart solution to help predict the prices of used devices that will enter the market in the future. For this we used past data of phones and tablets to build a machine learning model using simple ***linear regression.***

  Data was preprocessed and manually analyzed through EDA and treated for any inconsistencies. Statsmodel was used in python to build a model that fit the train data set and tested against the test data set. Then, the model was treated to meet all the linear regression assumptions.

  Listed below is the equation for out linear model using the coefficients of the parameters from the final model that predicts the prices of used devices:

```
Used Price = -0.4185421964635505 + 0.02157039452232516 * ( main_camera_mp ) +  0.014326716702493737 * ( selfie_camera_mp ) +
0.014542675108859027 * ( ram ) +  0.4204591474248068 * ( normalized_new_price ) +  0.4504573360614656 * ( weight_log ) +  -0.02
280952188765756 * ( years_since_release ) +  -0.09315350767078606 * ( os_Others ) +  0.043238778666174676 * ( 4g_yes )
```

Model Performance Summary

- Summary of steps and most important factors used by the ML model for prediction

I. We drop brand_name column and encode the other categorical data such as 4g, 5g & os.

II. Split the data into train and test data set in a 70:30 ratio, after splitting we got the following number of rows in train data = 2417 and the following number of rows in test data = 1037

III. The target variable is the normalized_used_price so we create new data frames ‘x’ and ‘y’ for independent and dependent variables separately.

IV. We build are preliminary regression model using stats model and we observe a large condition number that indicates high multicollinearity and high p-values suggesting insignificant features in the model.

V. We test all the other conditions required to assume linearity. Refer model assumption section.

VI. We review the results of the final model and compare the performance metrics such as R^2, Adj. R^2, Root mean squared error, mean absolute error and mean absolute percentage error.

Link to Appendix slide on model assumptions

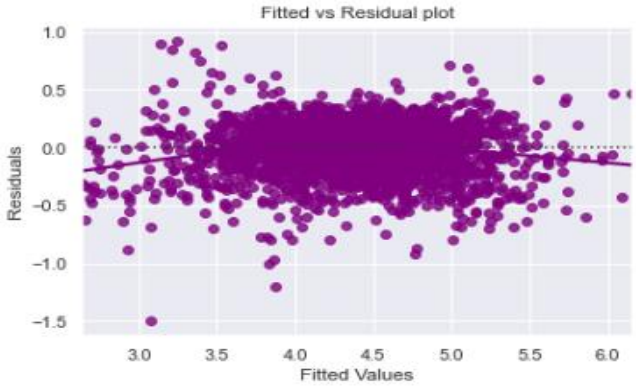
Model Performance Summary

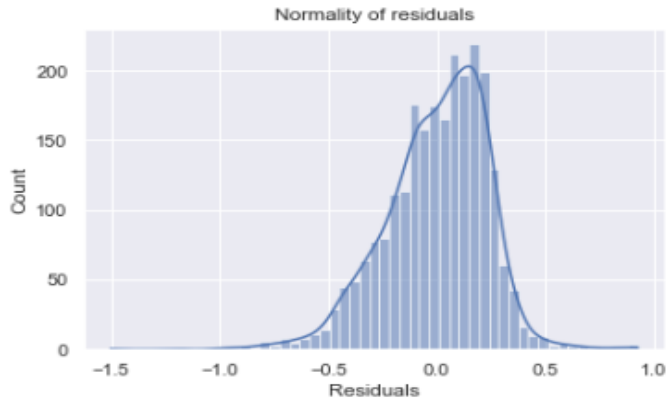
- Model Assumptions: We test for linear regression assumptions and fix them:
  a) Check for multicollinearity. Weight and screen_size were dropped. General Rule of thumb is:
     - 1 < VIF <= 5: There is low multicollinearity
     - 5 < VIF <= 10: There is moderate multicollinearity
     - VIF > 10: There is high multicollinearity
  b) Multiple column with p-values>0.05 (the agreed upon confidence interval) were dropped
  c) Then we test for linearity and independence of residuals.
  d) We also tested for normality by checking the distribution of residuals, by checking the Q-Q plot of residuals, and by using the Shapiro-Wilk test. `ShapiroResult(statistic=0.9655748009681702, pvalue=1.3800381585312504e-23)`

     Observations:
     - Since p-value < 0.05, the residuals are not normal as per the Shapiro-Wilk test.
     - However, based on the distribution plot and Q-Q plot, we can consider the distribution to be a good approximation of the normal distribution.
     - So, the assumption is satisfied.

Model Performance Summary

e) We tested for homoscedasticity by using the ***goldfeldquandt test***. We got a p-value greater than 0.05 hence, we can say that the residuals are homoscedastic. `[('F statistic', 1.0591342077114303), ('p-value', 0.159959816724501)]`





Model Performance Summary

- Summary of key performance metrics for training and test data in tabular format for comparison:

I. The train and test R^2 are 0.839 and 0.842, indicating that the model explains 83.9% and 84.2% of the total variation in the train and test sets respectively. Also, both scores are comparable.

II. RMSE values on the train and test sets are also comparable. This shows that the model is not overfitting.

III. MAE indicates that our current model is able to predict anime ratings within a mean error of 0.18 on the test set.

IV. MAPE of 4.47 on the test data means that we are able to predict within 4.47% of the used price.

OLS Regression Results

Dep. Variable:	normalized_used_price	R-squared:	0.840
Model:	OLS	Adj. R-squared:	0.839
Method:	Least Squares	F-statistic:	1579.
Date:	Fri, 13 May 2022	Prob (F-statistic):	0.00
Time:	17:16:45	Log-Likelihood:	85.528
No. Observations:	2417	AIC:	-153.1
Df Residuals:	2408	BIC:	-100.9
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.4185	0.082	-5.125	0.000	-0.579	-0.258
main_camera_mp	0.0216	0.001	15.757	0.000	0.019	0.024
selfie_camera_mp	0.0143	0.001	13.532	0.000	0.012	0.016
ram	0.0145	0.004	3.305	0.001	0.006	0.023
normalized_new_price	0.4205	0.011	39.175	0.000	0.399	0.442
weight_log	0.4505	0.016	28.710	0.000	0.420	0.481
years_since_release	-0.0228	0.003	-6.878	0.000	-0.029	-0.016
os_Others	-0.0932	0.027	-3.452	0.001	-0.146	-0.040
4g_yes	0.0432	0.015	2.914	0.004	0.014	0.072

Omnibus:	243.789	Durbin-Watson:	1.918
Prob(Omnibus):	0.000	Jarque-Bera (JB):	505.879
Skew:	-0.636	Prob(JB):	1.41e-110
Kurtosis:	4.846	Cond. No.	281.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE %
0	0.233558	0.181332	0.839888	0.839289	4.36079

Test Performance

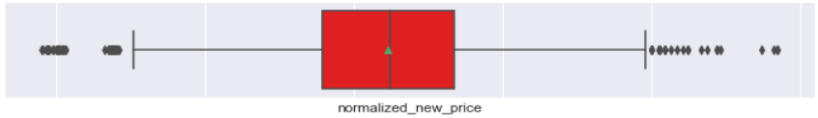
	RMSE	MAE	R-squared	Adj. R-squared	MAPE %
0	0.238242	0.183678	0.842632	0.841253	4.478225

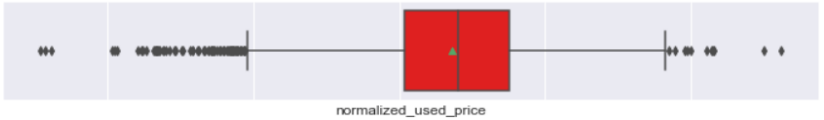
APPENDIX – A : EDA Results

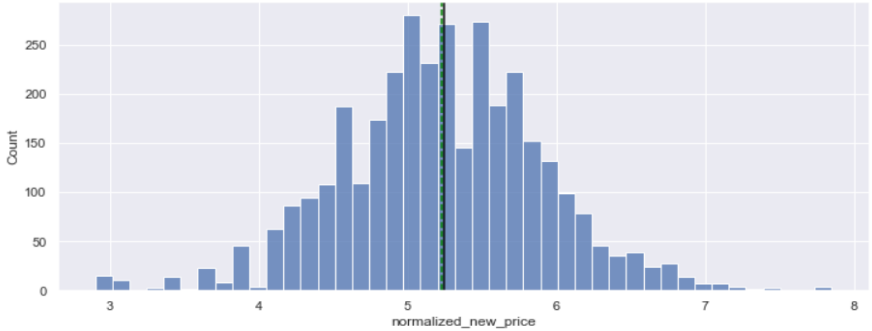
Statistical Summary of the data:

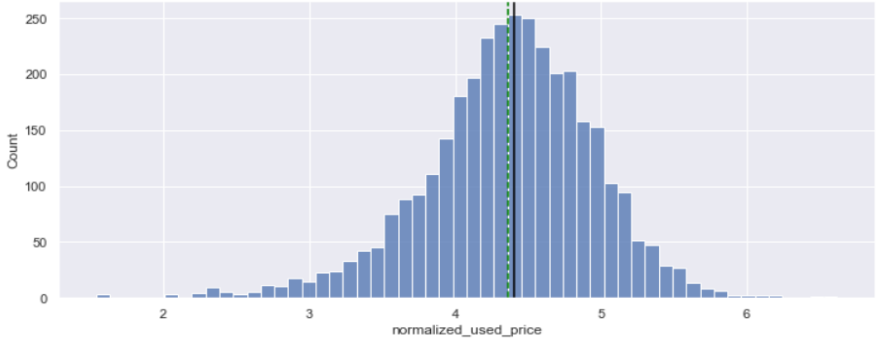
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
brand_name	3454	34	Others	502	NaN	NaN	NaN	NaN	NaN	NaN	NaN
os	3454	4	Android	3214	NaN	NaN	NaN	NaN	NaN	NaN	NaN
screen_size	3454.0	NaN	NaN	NaN	13.713115	3.80528	5.08	12.7	12.83	15.34	30.71
4g	3454	2	yes	2335	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5g	3454	2	no	3302	NaN	NaN	NaN	NaN	NaN	NaN	NaN
main_camera_mp	3275.0	NaN	NaN	NaN	9.460208	4.815461	0.08	5.0	8.0	13.0	48.0
selfie_camera_mp	3452.0	NaN	NaN	NaN	6.554229	6.970372	0.0	2.0	5.0	8.0	32.0
int_memory	3450.0	NaN	NaN	NaN	54.573099	84.972371	0.01	16.0	32.0	64.0	1024.0
ram	3450.0	NaN	NaN	NaN	4.036122	1.365105	0.02	4.0	4.0	4.0	12.0
battery	3448.0	NaN	NaN	NaN	3133.402697	1299.682844	500.0	2100.0	3000.0	4000.0	9720.0
weight	3447.0	NaN	NaN	NaN	182.751871	88.413228	69.0	142.0	160.0	185.0	855.0
release_year	3454.0	NaN	NaN	NaN	2015.965258	2.298455	2013.0	2014.0	2015.5	2018.0	2020.0
days_used	3454.0	NaN	NaN	NaN	674.869716	248.580166	91.0	533.5	690.5	868.75	1094.0
normalized_used_price	3454.0	NaN	NaN	NaN	4.364712	0.588914	1.536867	4.033931	4.405133	4.7557	6.619433
normalized_new_price	3454.0	NaN	NaN	NaN	5.233107	0.683637	2.901422	4.790342	5.245892	5.673718	7.847841

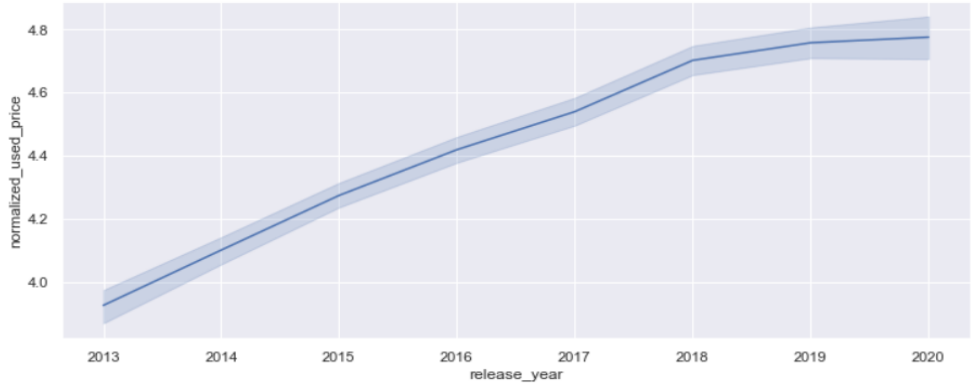
Distributions for normalized prices:



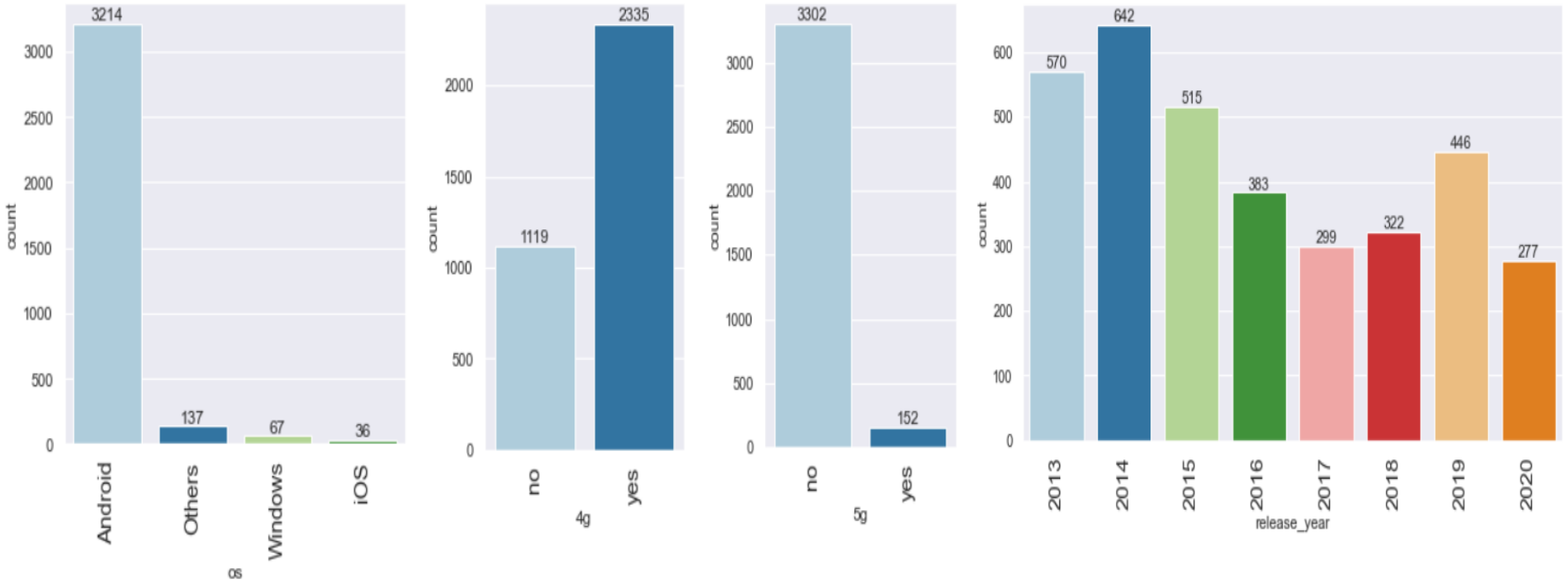






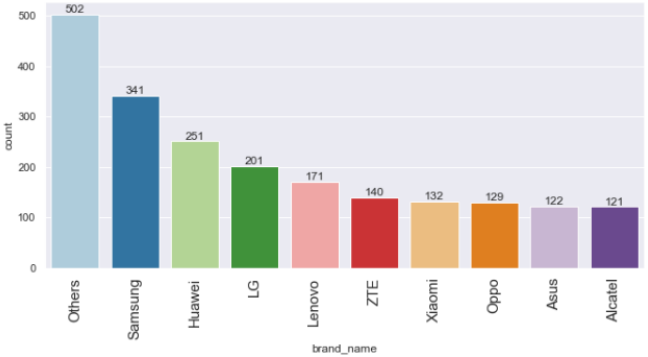


Distributions of counts for os, 4g, 5g and release year

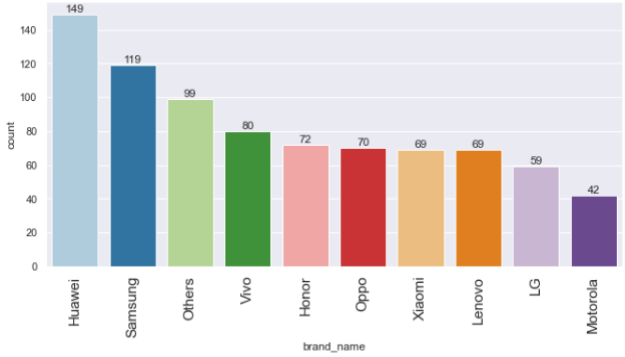


Brand-wise distribution (Top 10s)

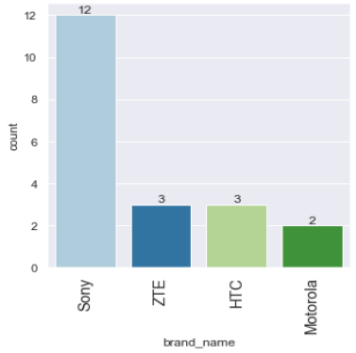
Brand name count plot



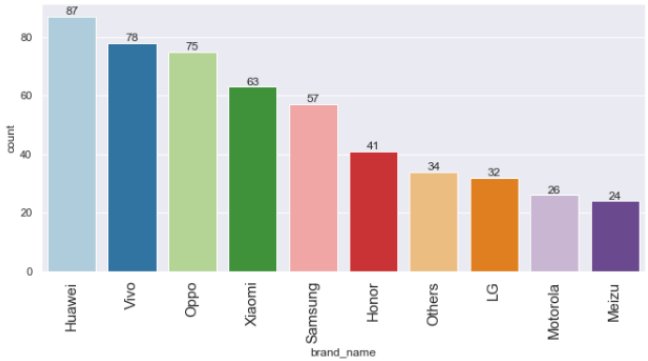
Brands with devices with large screens



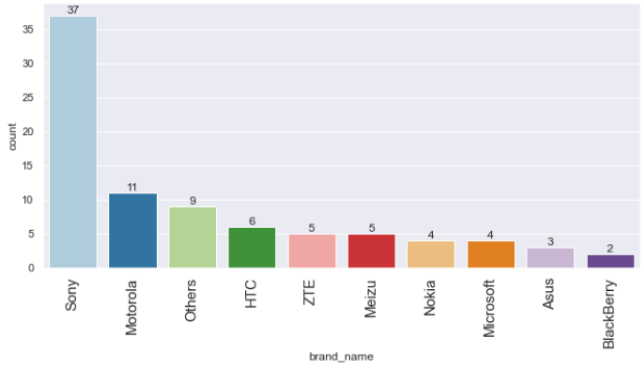
Brands with both good rear and front cameras



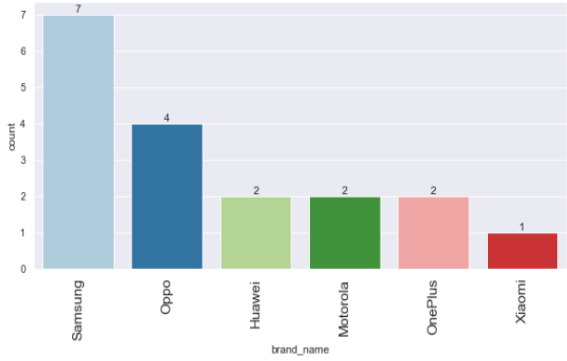
Brands with devices with good selfie cameras



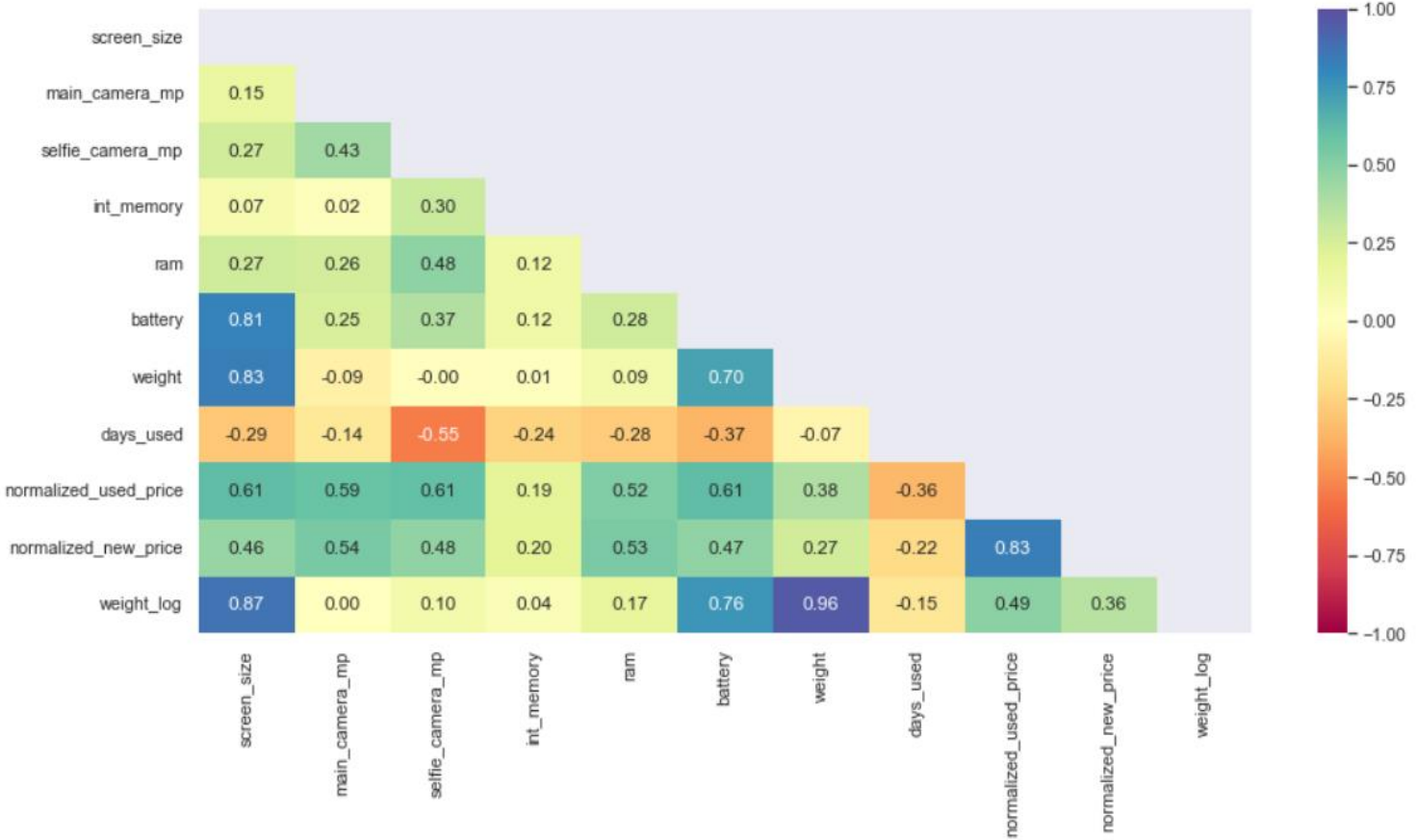
Brands with devices with good rear cameras



Brands with larger rams



Heat Map



APPENDIX – B : Model Performance Summary

First OLS Model performance results

OLS Regression Results

Dep. Variable:	normalized_used_price	R-squared:	0.844
Model:	OLS	Adj. R-squared:	0.843
Method:	Least Squares	F-statistic:	810.5
Date:	Fri, 13 May 2022	Prob (F-statistic):	0.00
Time:	17:16:42	Log-Likelihood:	115.68
No. Observations:	2417	AIC:	-197.4
Df Residuals:	2400	BIC:	-98.92
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.1123	0.259	-0.433	0.665	-0.621	0.396
screen_size	0.0232	0.003	6.977	0.000	0.017	0.030
main_camera_mp	0.0210	0.001	14.955	0.000	0.018	0.024
selfie_camera_mp	0.0139	0.001	12.958	0.000	0.012	0.016
int_memory	9.873e-05	6.65e-05	1.485	0.138	-3.17e-05	0.000
ram	0.0191	0.005	3.853	0.000	0.009	0.029
battery	-1.655e-05	7.07e-06	-2.341	0.019	-3.04e-05	-2.69e-06
weight	-0.0002	0.000	-0.834	0.404	-0.001	0.000
days_used	5.303e-05	3.03e-05	1.750	0.080	-6.39e-06	0.000
normalized_new_price	0.4163	0.011	36.960	0.000	0.394	0.438
weight_log	0.3404	0.058	5.842	0.000	0.226	0.455
years_since_release	-0.0234	0.004	-5.290	0.000	-0.032	-0.015
os_Others	-0.0167	0.030	-0.562	0.574	-0.075	0.042
os_Windows	0.0349	0.036	0.968	0.333	-0.036	0.106
os_iOS	-0.0517	0.044	-1.165	0.244	-0.139	0.035
4g_yes	0.0456	0.015	3.015	0.003	0.016	0.075
5g_yes	-0.0501	0.031	-1.636	0.102	-0.110	0.010

Omnibus:	230.772	Durbin-Watson:	1.913
Prob(Omnibus):	0.000	Jarque-Bera (JB):	466.370
Skew:	-0.615	Prob(JB):	5.36e-102
Kurtosis:	4.766	Cond. No.	1.95e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.95e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE %
0	0.230663	0.179806	0.843833	0.842726	4.31857

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE %
0	0.234764	0.181291	0.847193	0.844644	4.417644

Great Learning

Happy Learning !