University of Austin Texas
PG-DSBA Certificate Program
Presenter: Uchenna Nwosu

ReCell Presentation

Contents

Problem Overview and Objective

Business Description

IDC (International Data Corporation) predicts that the used phone market would be worth $52.7bn by 2023. This growth can be attributed to an uptick in demand for used smartphones that offer considerable savings compared to new ones. Refurbished and used devices are cost-effective alternatives. Maximizing the longevity of mobile-phones through second-hand trade is being environmentally savvy.

Objective

ReCell, a startup, aims to develop an ML-based solution for dynamic pricing of used and refurbished smartphones. It wants to analyze the available data and build a linear regression model that can predict the price of a used phone. It also seeks to identify factors that significantly influence price.

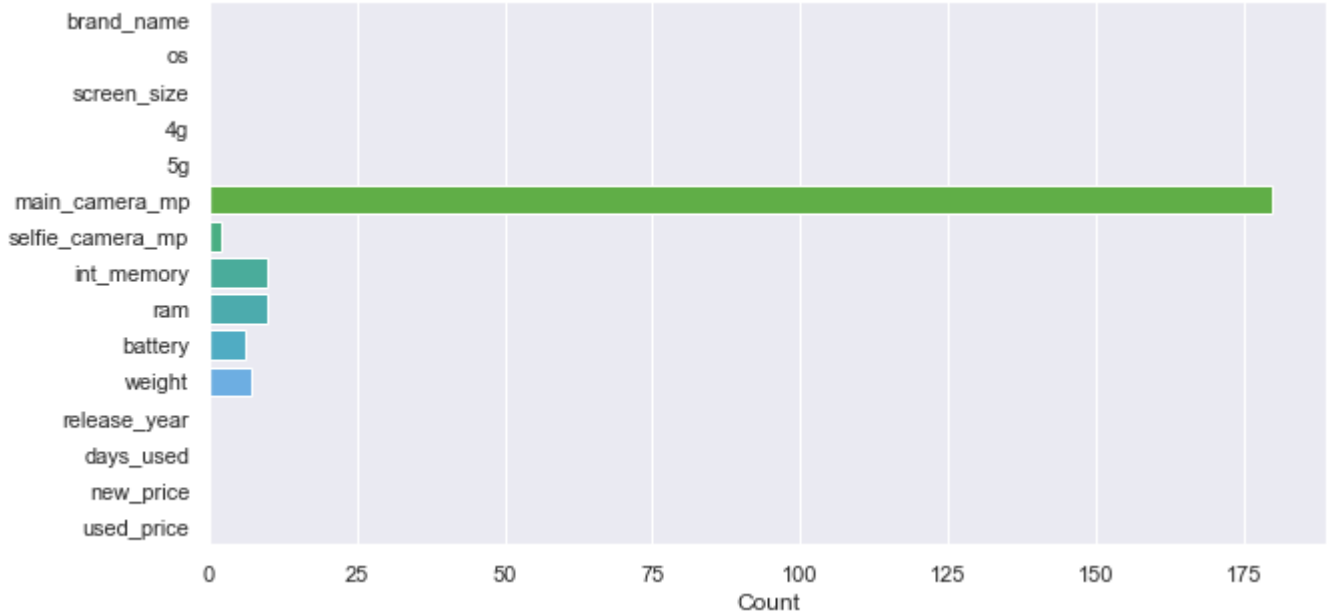
Data Overview

Feature name	Type
brand name	category
os	category
screen size	numeric
4g	category
5g	category
main camera (mp)	numeric
selfie camera (mp)	numeric
internal memory	numeric
ram	numeric
battery	numeric
weight	numeric
release year	numeric
days used	numeric
new price	numeric
used price	numeric

Sample size provided:

3571 rows X 15 columns

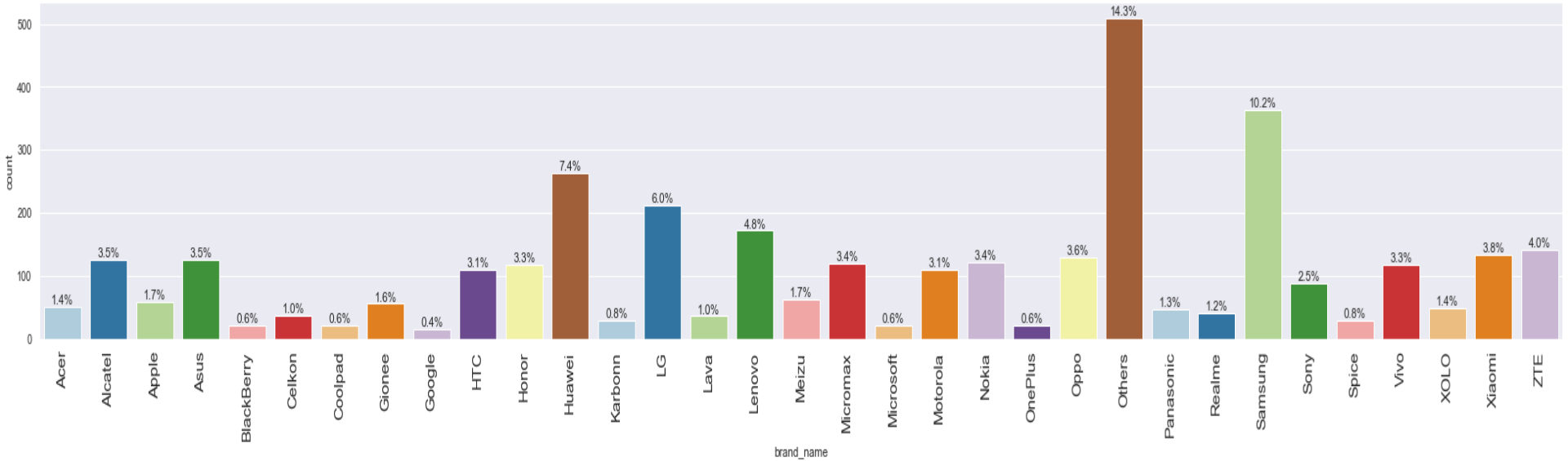
EDA: for missing value analysis



Observations:

- **Almost all the missing values are from one feature**
- **Only numeric features are missing values**

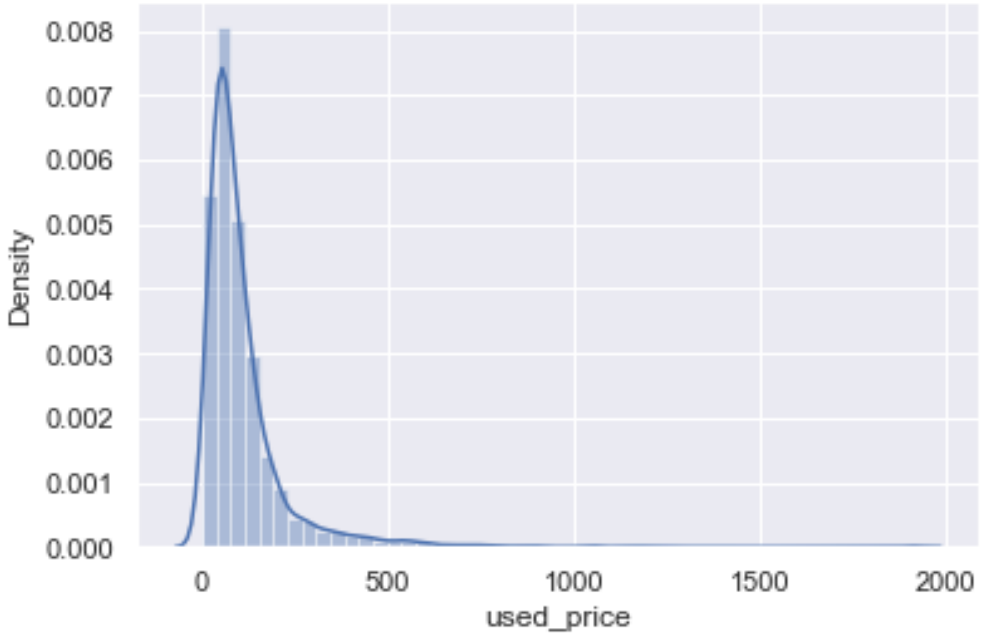
EDA: Phone brand distribution in the sample



Observations:

- The sample looks randomly distributed
- Samsung is the highest occurring named brand as expected

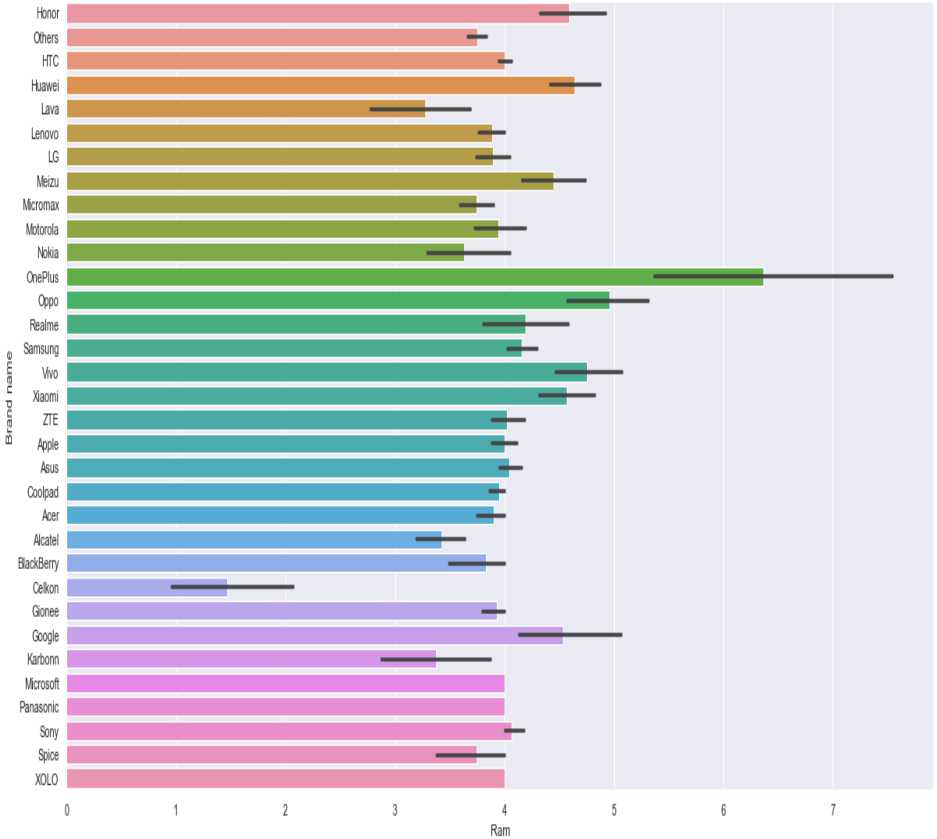
EDA - for regression modelling

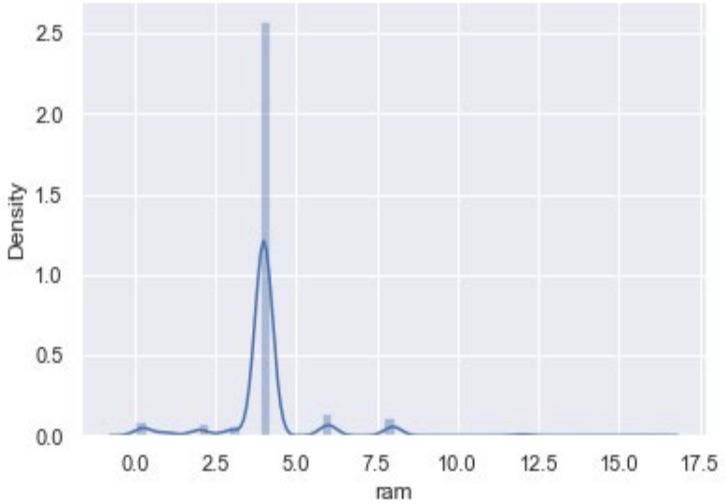


Observations:

- The distribution for used phone prices is significantly right skewed.

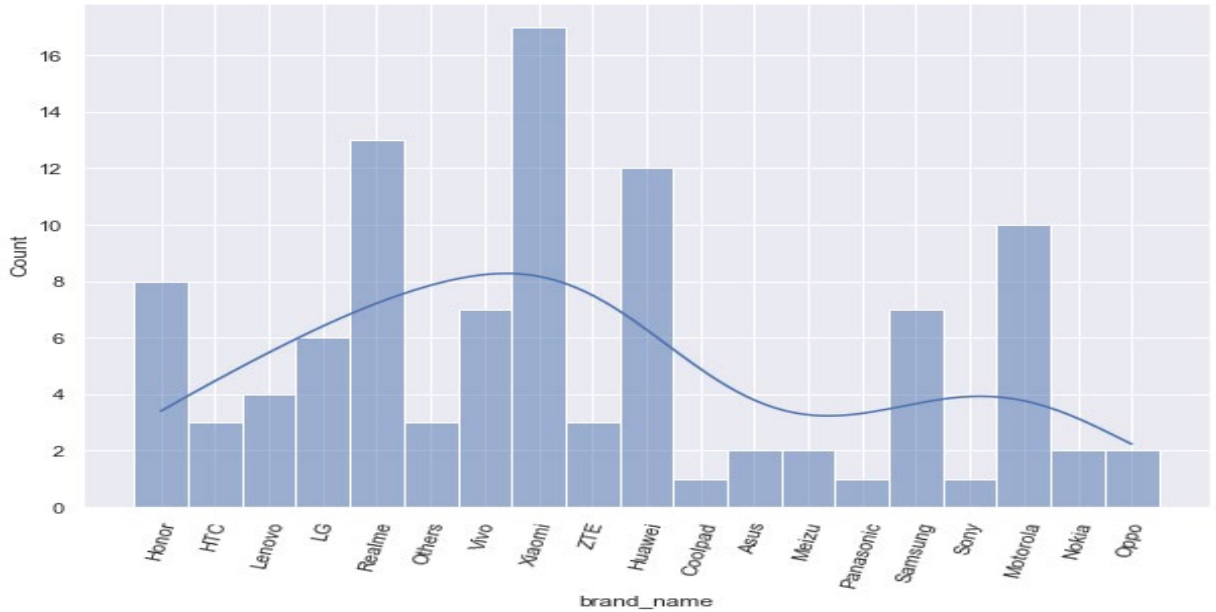
Amount of Ram by phone brand





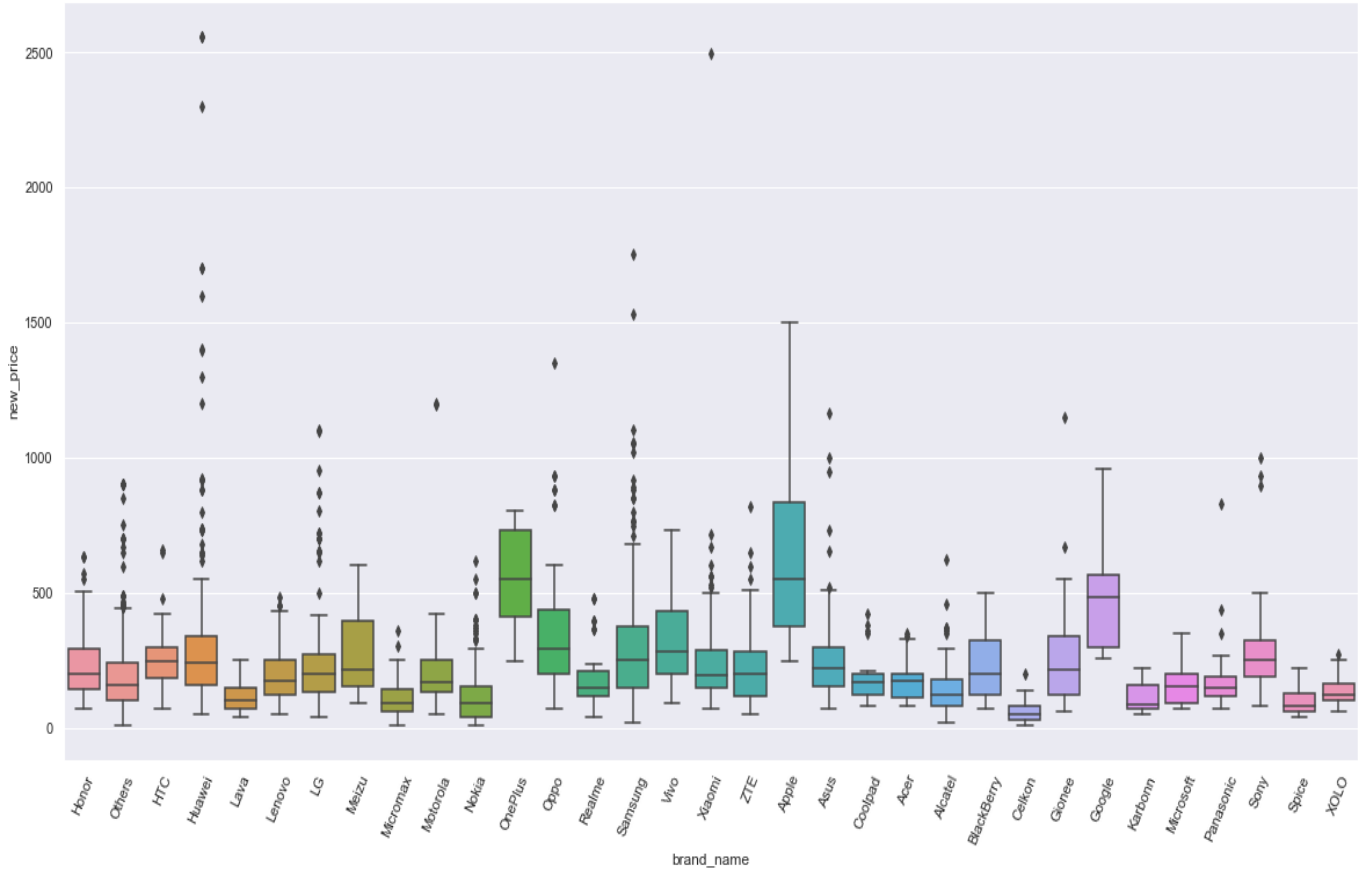
The attempt to treat outliers in the Ram feature changed the distribution to an even distribution that yielded poor results. The outlier treatment was made to bypass the Ram data thereafter.

EDA – general insights



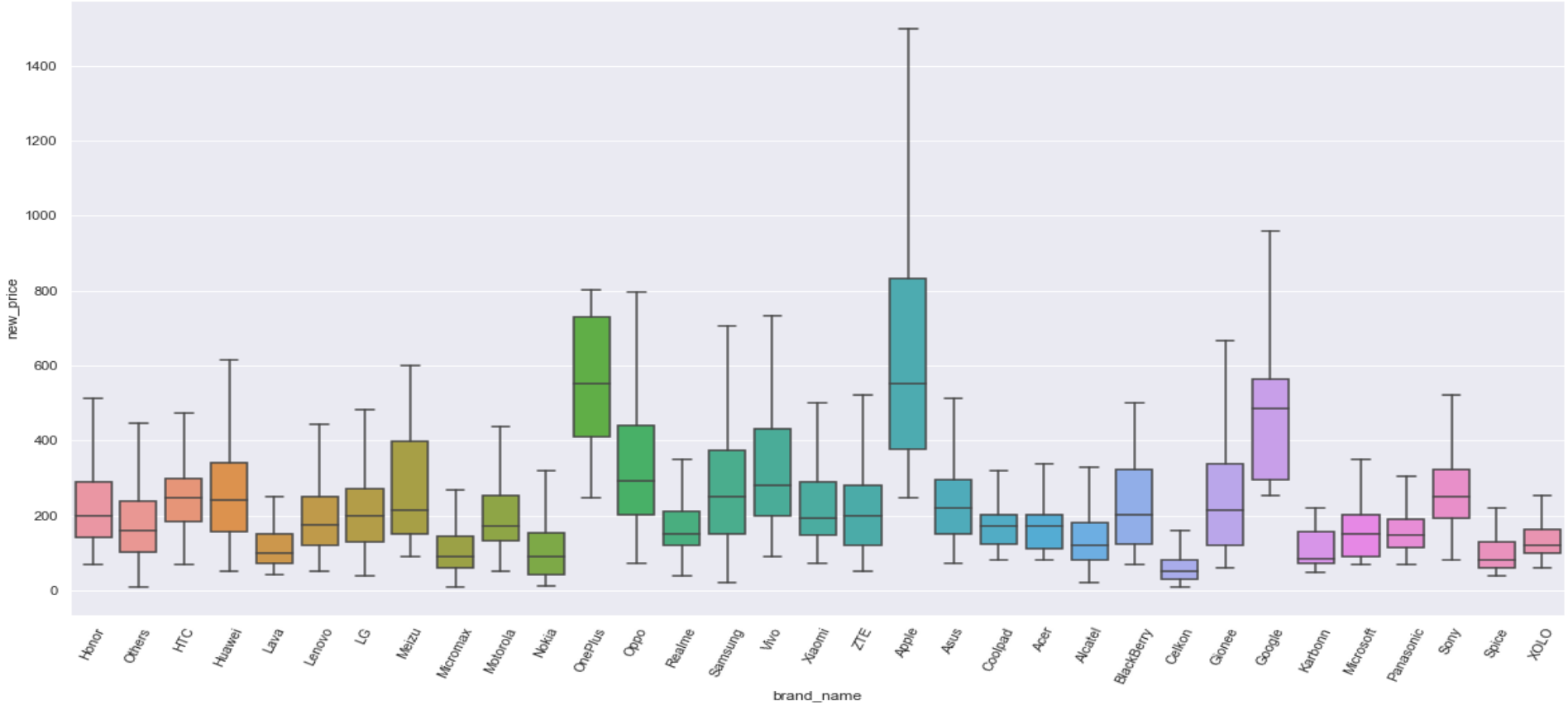
- **Distribution of budget phones that while costing not more than the median price of roughly 190 dollars, still offer 8 mega pixel selfie cameras**

EDA – for outlier detection



The boxplots show potential new price outlier values that could negatively impact the model.

EDA: showing new price after outlier treatment

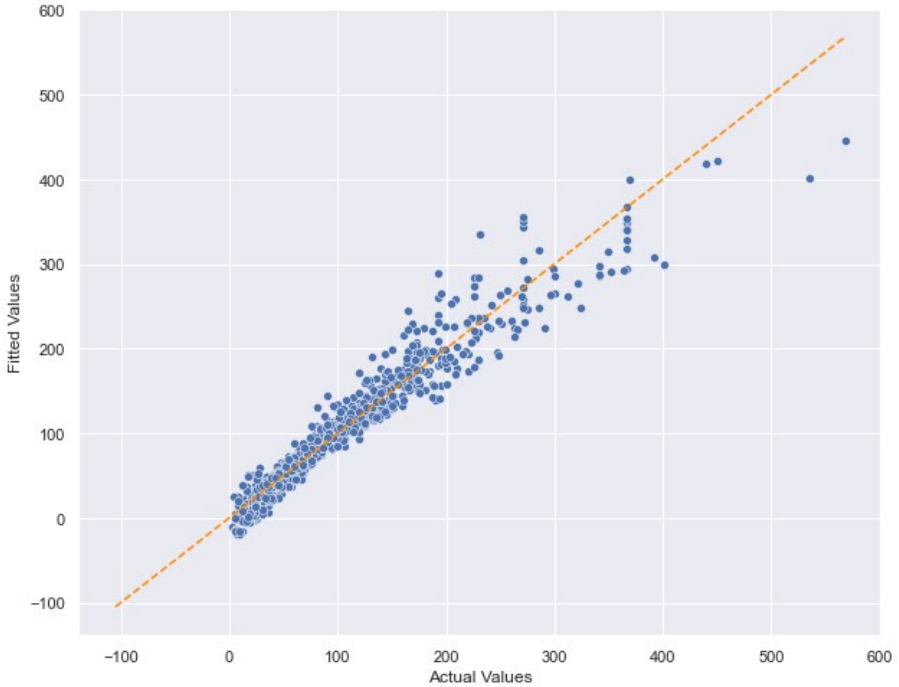


EDA – showing correlations between numeric features

Strong linear correlations between predictor variables like new price, and the response variable, used price, are ground assumptions of linear regression modelling. However, strong correlations among the predictors is undesirable because they lead to variance inflation measured by the variance inflation factor - VIF

Multiple linear regression MLS prior to optimization

Actual Values	Fitted Values	Residuals
40.14	34.642094	5.497906
6.16	-14.182414	20.342414
39.44	37.933512	1.506488
76.07	82.197479	-6.127479
29.85	5.028945	24.821055
262.68	224.449620	38.230380
91.47	98.958202	-7.488202
36.59	38.406706	-1.816706
33.13	26.238476	6.891524
223.76	236.692533	-12.932533
36.09	14.492023	21.597977
67.70	76.805151	-9.105151
131.34	126.366847	4.973153
64.50	62.707683	1.792317
79.43	71.520492	7.909508
54.48	57.356822	-2.876822
57.24	56.094754	1.145246
90.85	95.273160	-4.423160
67.60	78.268156	-10.668156
69.38	68.163608	1.216392

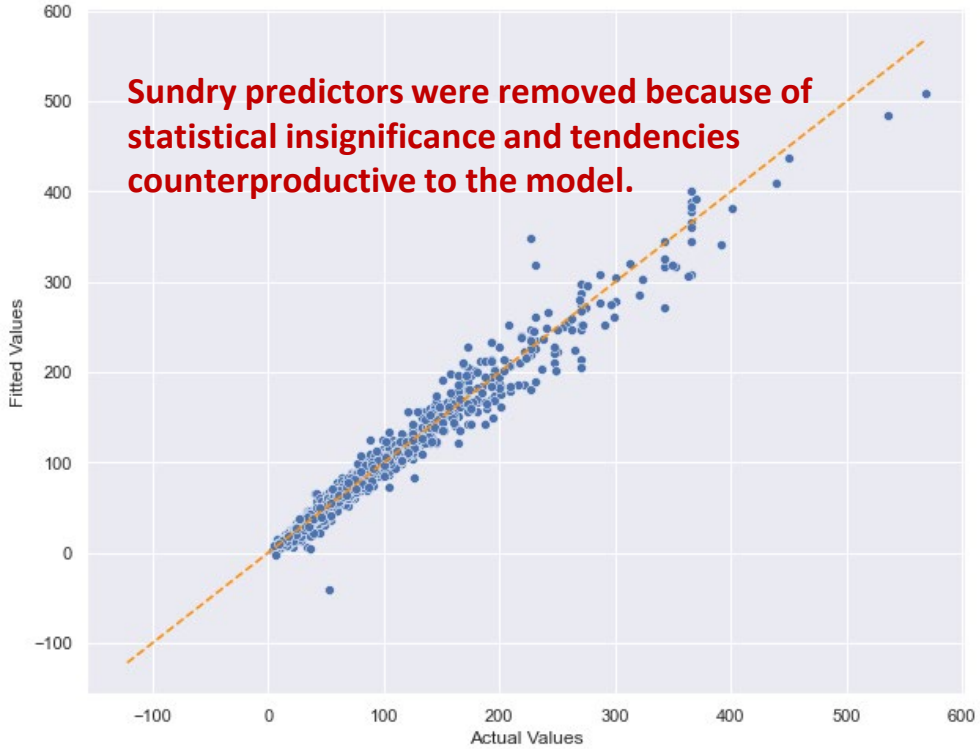


There is much error in the real to predicted values. When they converge, they align with the orange line.

RMSE	MAE	R-squared	Adj. R-squared	MAPE
19.231338	12.623563	0.934176	0.931415	20.68172

MLS - with optimization by polynomial features

Actual Values	Fitted Values	Residuals
27.10	33.982504	-6.882504
30.45	27.495897	2.954103
45.07	50.132289	-5.062289
108.04	107.405529	0.634471
80.44	68.879641	11.560359
60.59	56.563053	4.026947
17.65	15.113922	2.536078
351.63	287.040476	64.589524
124.83	104.741477	20.088523
187.96	197.083213	-9.123213
60.32	67.949325	-7.629325
34.91	36.669937	-1.759937
28.59	32.114572	-3.524572
45.43	47.730445	-2.300445
27.75	32.305656	-4.555656



RMSE	MAE	R-squared	Adj. R-squared	MAPE
14.311925	8.744904	0.963545	0.959528	10.673887

Insights and Recommendations

- Many features were found unworthy of being predictors. They were removed from the model because they were either statistically insignificant or counter productive.
- The original or used price accounts for close to 70% of the linear variation. The days used is a fairly significant predictor and has a negative correlation to the used price which is the target variable.
- The phone brand can be a very significant contributor to the model: Apple, Google, OnePlus, and CelKon on a standardized scale contributed more positively to the regression. They are bound to command a premium price. This is significant considering the sample had less than 2% Apple phone data.
- The volume and quality of the data is critical to model formation because linear models thrive on normally distributed data. To this end, as much as possible effort should be intensified towards comprehensive data gathering.
- The 4g phones have a negative correlation with price. More test modeling is required to isolate the cause

The model developed should be used as a framework for pricing structure. It certainly in most cases gives a reasonable and consistent prediction.

The final features that constitute the model

	Coefficients
screen_size	0.392441
main_camera_mp	-0.389673
selfie_camera_mp	0.685559
int_memory	0.139785
ram	1.030857
days_used	-0.087758
new_price	0.403464
brand_name_Apple	31.579644
brand_name_Celkon	-17.916367
brand_name_Google	27.697592
brand_name_OnePlus	36.071226
os_Others	1.523852
4g_yes	-5.230707
5g_yes	23.507918
Intercept	52.848395

greatlearning
Power Ahead

Happy Learning !