ReCell

USED CELL PHONE PRICE PREDICTION

Objective

The rising potential of this comparatively under-the-radar market fuels the need for an ML-based solution to develop a dynamic pricing strategy for used and refurbished smartphones. ReCell, a startup aiming to tap the potential in this market, has hired you as a data scientist. They want you to analyze the data provided and build a linear regression model to predict the price of a used phone and identify factors that significantly influence it.

Context

Buying and selling used smartphones used to be something that happened on a handful of online marketplace sites. But the used and refurbished phone market has grown considerably over the past decade, and a new IDC (International Data Corporation) forecast predicts that the used phone market would be worth $52.7bn by 2023 with a compound annual growth rate (CAGR) of 13.6% from 2018 to 2023. This growth can be attributed to an uptick in demand for used smartphones that offer considerable savings compared with new models.

Refurbished and used devices continue to provide cost-effective alternatives to both consumers and businesses that are looking to save money when purchasing a smartphone. There are plenty of other benefits associated with the used smartphone market. Used and refurbished devices can be sold with warranties and can also be insured with proof of purchase. Third-party vendors/platforms, such as Verizon, Amazon, etc., provide attractive offers to customers for refurbished smartphones. Maximizing the longevity of mobile phones through second-hand trade also reduces their environmental impact and helps in recycling and reducing waste. The impact of the COVID-19 outbreak may further boost the cheaper refurbished smartphone segment, as consumers cut back on discretionary spending and buy phones only for immediate needs.

Data Overview: The dataset file contains the used phone data with following specifications:

Variable name	Data types	Description	Missing values	Unique values
1. brand_name:	categorical	Name of manufacturing brand	0	34
2. os:	categorical	OS on which the phone runs	0	4
3. screen_size:	numeric	Size of the screen in cm	0	127
4. 4g:	categorical	Whether 4G is available or not	0	2
5. 5g:	categorical	Whether 5G is available or not	0	2
6. main_camera_mp:	numeric	Resolution of the rear camera in megapixels	180	44
7. selfie_camera_mp:	numeric	Resolution of the front camera in megapixels	2	37
8. int_memory:	numeric	Amount of internal memory (ROM) in GB	10	16
9. ram:	numeric	Amount of RAM in GB	10	14
10. battery:	numeric	Energy capacity of the phone battery in mAh	6	354
12. weight:	numeric	Weight of the phone in grams	7	613
12. release_year:	numeric	Year when the phone model was released	0	8
13. days_used:	numeric	Number of days the used/refurbished phone has been used	0	930
14. new_price:	numeric	Price of a new phone of the same model in euros	0	3099
15. used_price:	numeric	Price of the used/refurbished phone in euros	0	3044

DATA OVERVIEW: MISSING VALUES TREATMENT

- Brief description of significant manipulations made to raw data

Observations	Variables	Missing	Dependent variable
3571	15	215	used_price

Variable name	Missing data description	Treatment
main_camera_mp: selfie_camera_mp: battery weight:	205 cells have missing values which didn't show any pattern.	Mean of these fields were imputed in the empty cells.
int_memory, ram: battery	Block of 20 values were missing, from the same rows in columns int_memory and ram and 2 were missing in battery as well.	Last Observation Carried Forward (LOCF)were imputed as it was a block of 10 rows with same brand and similar specifications. Next Observation Carried Backward (NOCB)were imputed in battery in these rows

DATA OVERVIEW

- Brief description of significant manipulations made to raw data

Observations	Variables	Missing	Dependent variable
3571	15	215	used_price

Categorical Variables: Encoding into numeric.	
Variable name	**Treatment**
brand_name	Frequency encoding was used to as 34 unique values were present.
4g, 5g, os	One Hot Encoding was used as there were fewer unique values.

Variable name	Feature Engineering
screen_size	Converted from cm to inches

DATA OVERVIEW: OUTLIERS TREATMENT & FEATURE ENGINEERING

- Brief description of significant manipulations made to raw data

Variable name	Outliers detection and treatment	Standardization of variable
• int_memory, • main_camera_mp: • selfie_camera_mp: • battery • Weight • new_price • used_price	• IQR was used to detect outliers in all the numeric fields. • Outliers in the data were treated by flooring and capping.	• StandardScaler of sklearn library was used on all independent numeric variables to standardise all the variables to similar scale. • Normalization was done by using MinMaxScaler of sklearn library, so that distributions would be more Gaussian shaped.
• Screen_size • Weight • ram:	• Screen size had very big values so the data may be erroneous so they were removed. • Weight had very big values so the data may be erroneous so they were removed. • ram had majority values around 4 so outliers >8 were removed.	

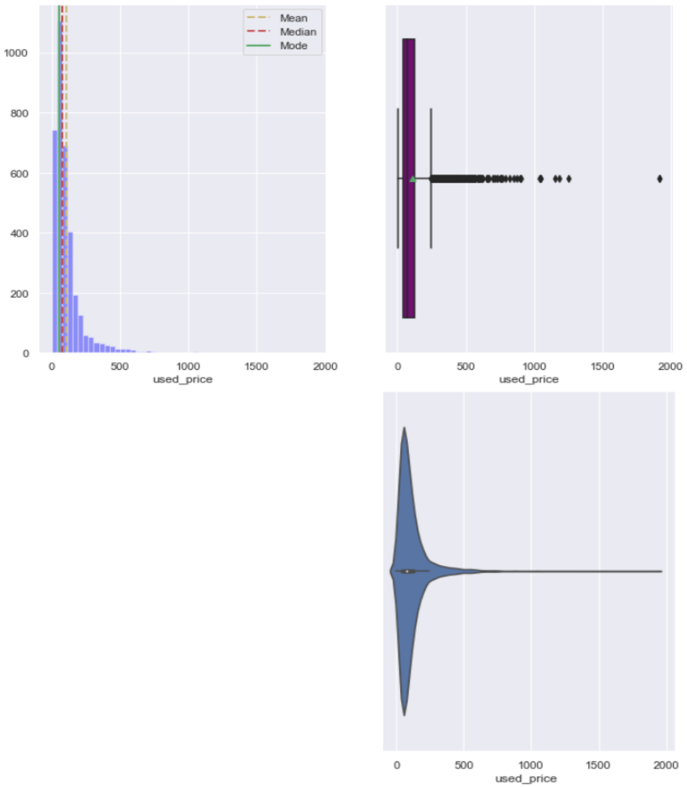
EXPLORATORY DATA ANALYSIS

- Graphs and observation about the target attribute:

Observations

- It is the target variable.
- The distribution of used_price is heavily skewed to the right.
- The outliers to the right indicate that many cell phones, though used, have a very high Prices.
- Mean is 109.9 much greater than median due to extreme values towards higher end.
- There are a lot of outliers towards right, larger side.Majority of values 50% are between 45 and 126

SPREAD OF DATA FOR USED_PRICE



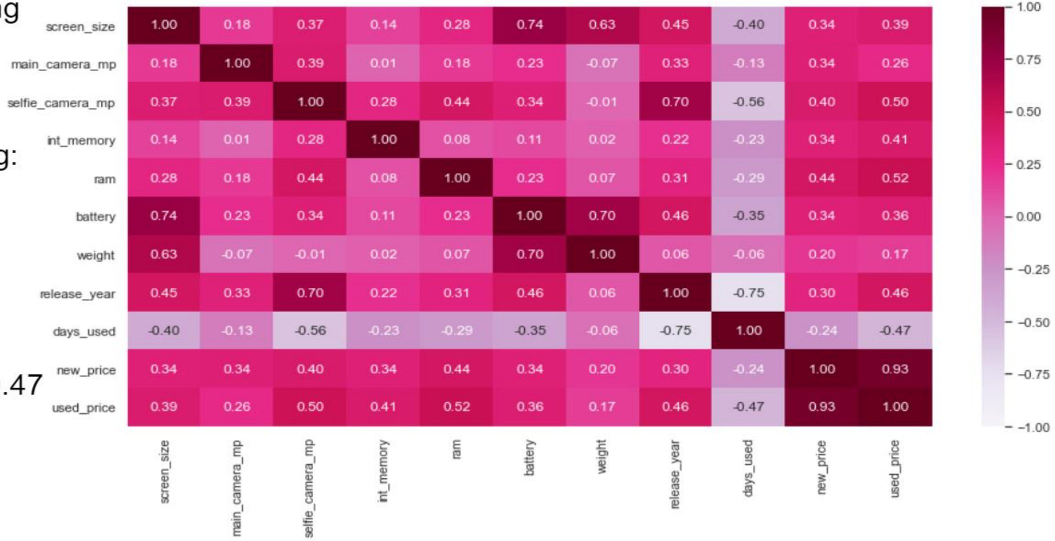
EDA

- Graphs showing the factors most heavily impacting the target attribute

Observations:

used_price has significant correlation with the following:

- high positive correlation with new_price 0.93
- Positive Correlation with ram 0.52
- Positive Correlation with selfie_camera_mp: 0.50
- somewhat negative correlation with days_used: -0.47

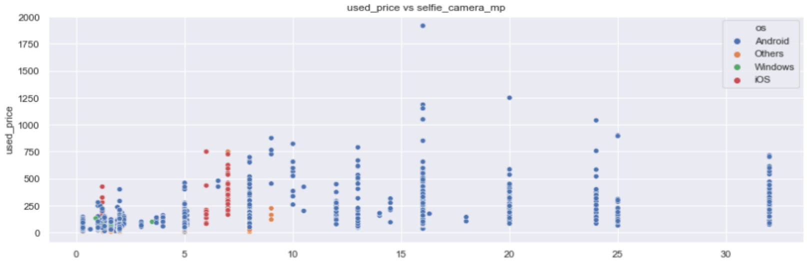


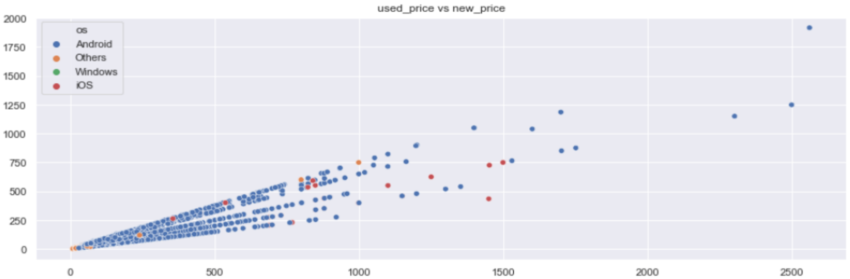
These fields also have some correlation:

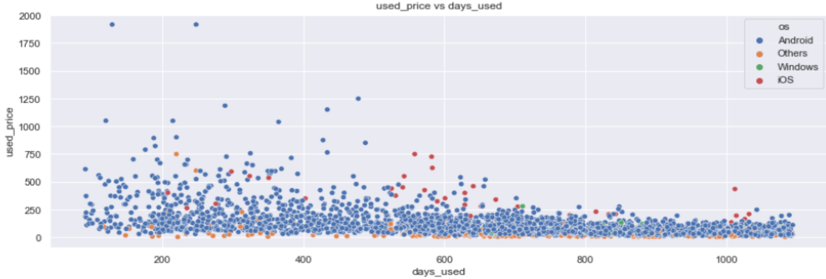
- release_year and selfie_camera_mp have positive correlation: 0.7
- weight and screen_size: 0.63,
- weight and battery: 0.7
- battery and screen size also have positive high correlation: 0.74
- days_used has high negative correlation with selfie_camera_mp: -0.56

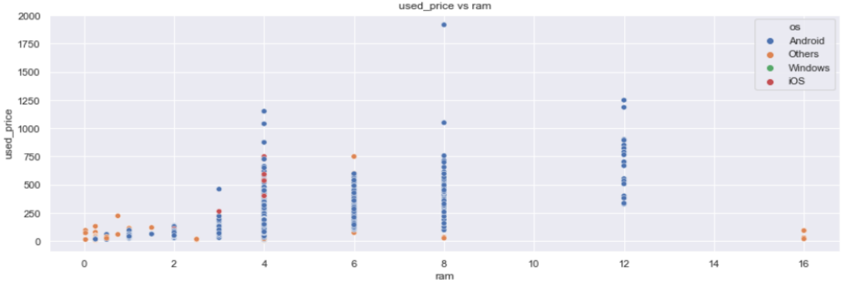
EDA

- Graphs showing the factors most heavily impacting the target attribute





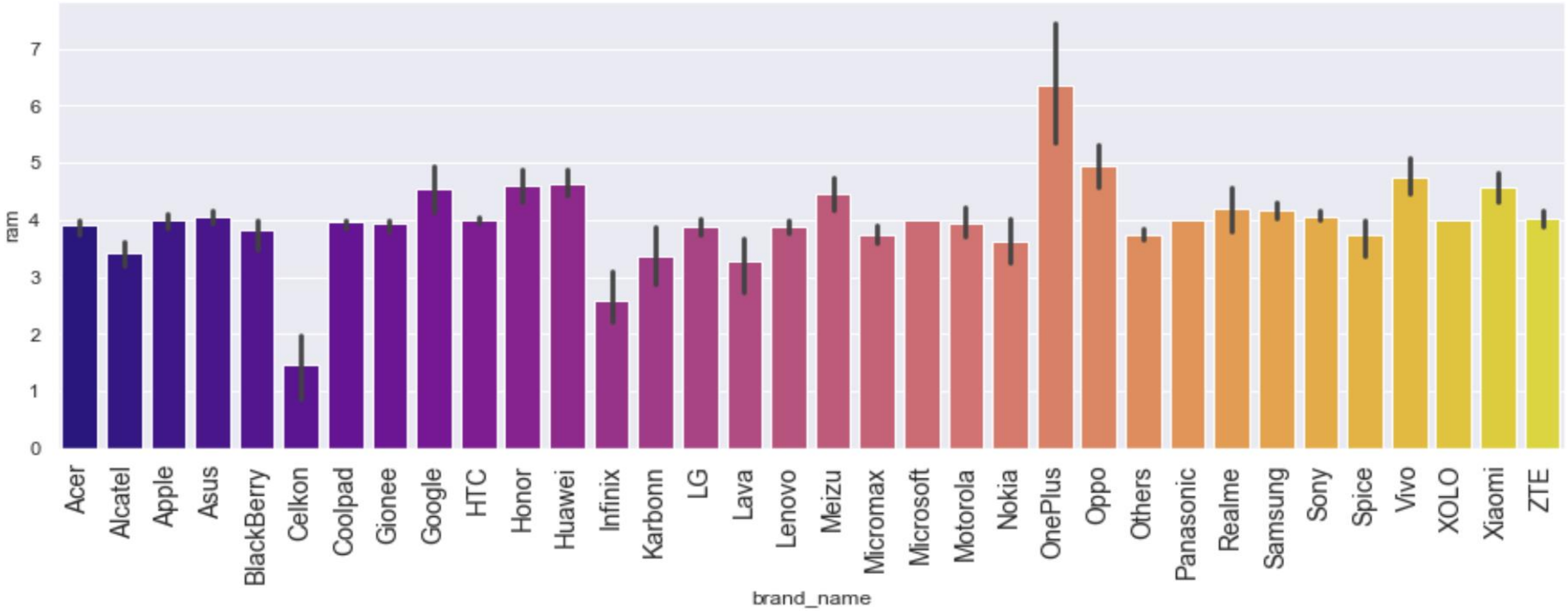




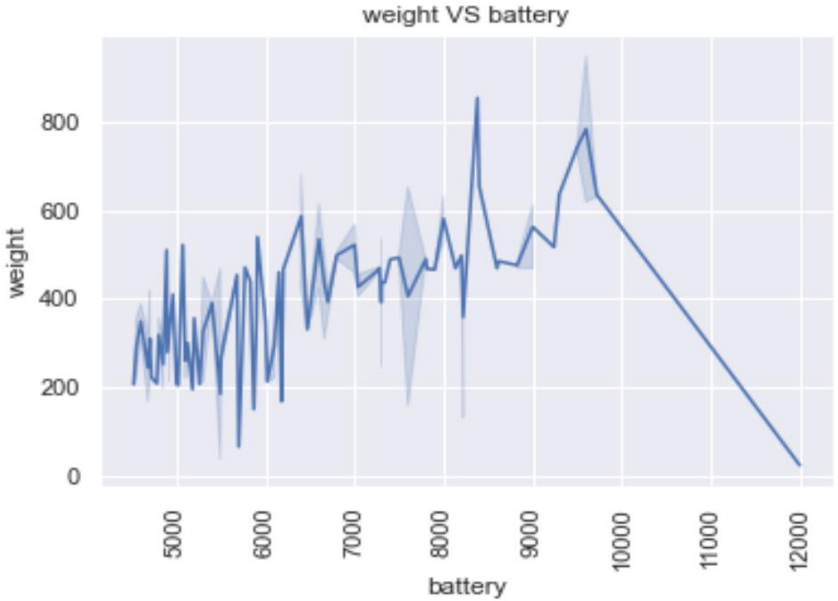
- Used_price increases if new_price is higher
- Used_price increases if ram is higher
- Used_price increases if selfie_camera_mp is better
- Used_price decreases is days_used is higher

HOW DOES THE AMOUNT OF RAM VARY WITH THE BRAND?

- brands vary in ram in range 1.5-6GB or 4gb ram on average



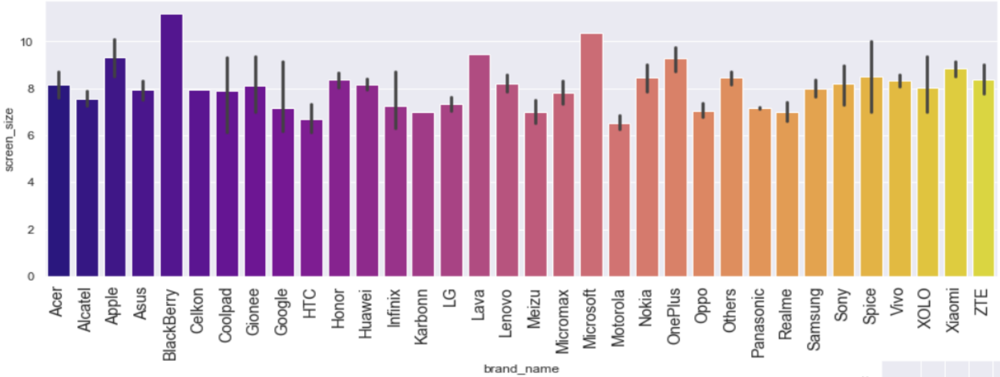
HOW DOES THE WEIGHT VARY FOR PHONES OFFERING LARGE BATTERIES (MORE THAN 4500 MAH)?

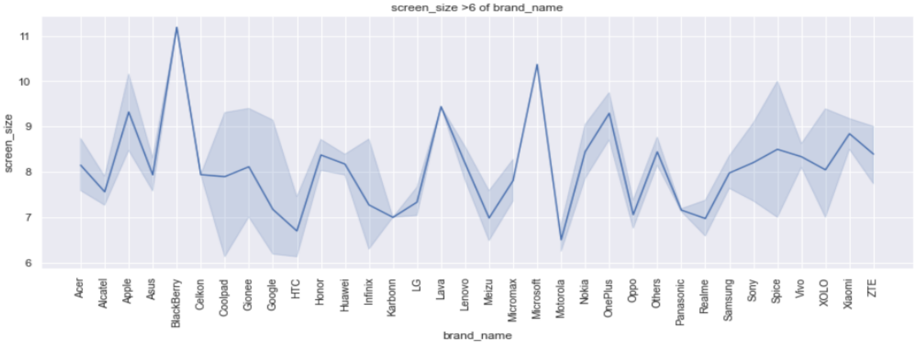


- On average phones with higher battery also have higher weight.
- Some phones have weight too high to be a cell phone weight. Cell phone heavier than 600 grams would be too heavy, maybe their units are wrong or data is erroneous.

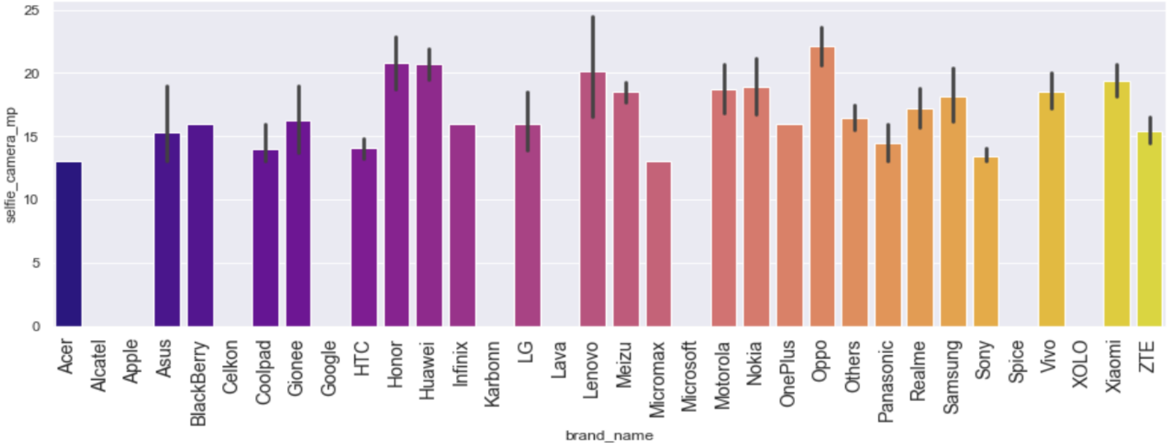
HOW MANY PHONES ARE AVAILABLE ACROSS DIFFERENT BRANDS WITH A SCREEN SIZE LARGER THAN 6 INCHES?

- Screen size on average are between 6"and 8".
- There are also some very big screen size present, like blackberry 11"+ and Microsoft 10"+, which may be some error in the data.

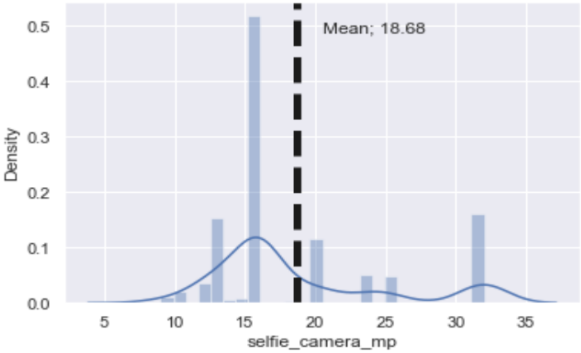




WHAT IS THE DISTRIBUTION OF BUDGET PHONES OFFERING GREATER THAN 8MP SELFIE CAMERAS ACROSS BRANDS?



- Specifications of selfie_camera_mp > 8mp
- max is 32mp
- mean is 18.7mp
- 50% of these phones have under 16mp



MODEL PERFORMANCE SUMMARY

Overview of ML model and its parameters:

- Multiple Linear Regression model was built to
- find dependency of target variable: used_price on predictors and
- predict fitted values and compare them to actual values
- Total row and column after data preprocessing:
- Target variable: used_price
- Predictors: "screen_size",'main_camera_mp','selfie_camera_mp','int_memory','ram','battery','weight','release_year',days_used',new_price',brand_name_encode',4g_yes','5g_no','os_Android','os_Others','os_iOS'
- Data was divided into Train and Test at 70:30 ratio.
- Number of rows in train data = 2499
- Number of rows in test data = 1072

MODEL PERFORMANCE SUMMARY

The first ML Performance was tested using the following testing metrics

Training Performance

RMSE	*MAE*	*R2*	*adjR2*	*MAPE*
0.082899	0.06852	0.990231	0.990039	1.697228

Testing Performance

RMSE	*MAE*	*R2*	*adjR2*	*MAPE*
0.082181	0.070285	0.989854	0.98937	1.699362

Observations

- The testing R2 is 98.9%, indicating that the model explains 98.9% of the variation in the train data. So, the model is not underfitting.

- MAE and RMSE on the train and test sets are comparable, which shows the model is not overfitting.
- MAE indicates that our current model is able to predict used phone prices within a mean error of 0.070 on the test data.
- MAPE on the test set suggests we can predict within 1.7% of the used phone prices.

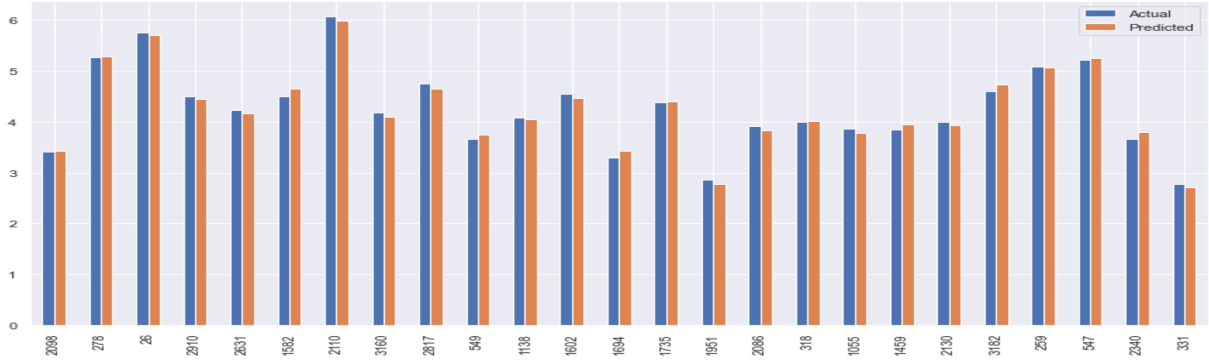
ASSUMPTIONS OF LINEAR REGRESSION

We will be checking ML model on linear regression assumptions.

1. No multicollinearity : Final ML model features did not have vif > 5

	features	vif
0	const	3547358.20
1	selfie_camera_mp	2.86
2	int_memory	1.29
3	Ram	1.80
4	days_used	2.62
5	new_price	2.98
6	4g_yes	2.40
7	os_iOS	1.70

FINAL MODEL PERFORMANCE SUMMARY (Actual vs Predicted Used Price)



	Actual values	predicted values
2098	3.418382	3.422896
278	5.276430	5.280973
26	5.751493	5.704627
2910	4.499476	4.458647
2631	4.237001	4.165908
1582	4.495132	4.643285
2110	6.067916	5.992102
3160	4.179604	4.096506
2817	4.751605	4.648188
549	3.670970	3.755021

Observations

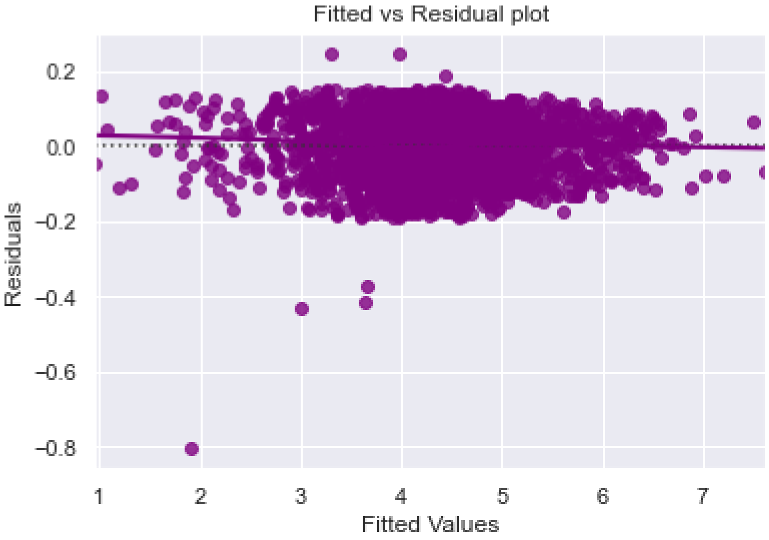
- We can observe here that our model has returned pretty good prediction results, and the actual and predicted values are comparable

- We can also visualize comparison result as a bar graph.

- Note: As the number of records is large, for representation purpose, we are taking a sample of 25 records only

ASSUMPTIONS OF LINEAR REGRESSION

2.Linearity of variable and independence of error terms

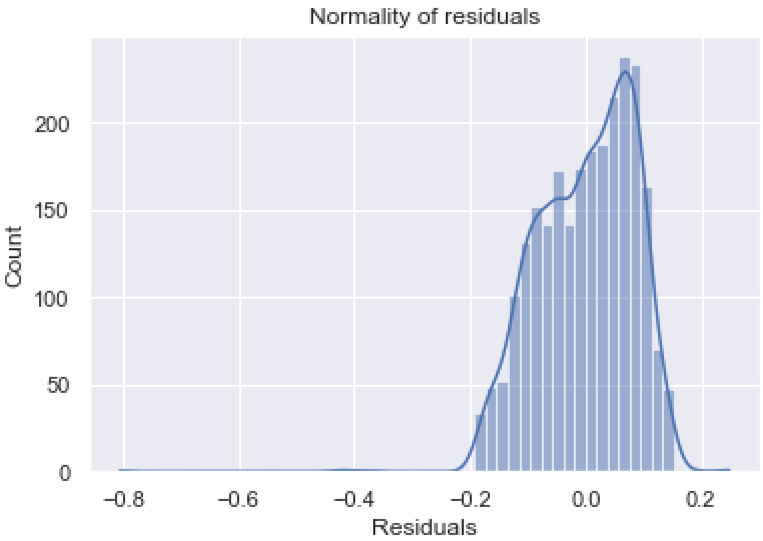


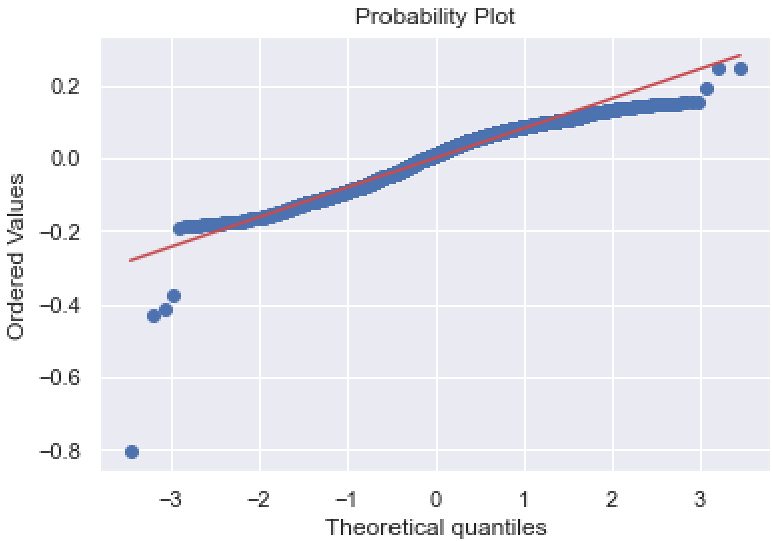
Observations

- The scatter plot shows the distribution of residuals (errors) vs fitted values (predicted values).
- If there exist any pattern in this plot, we consider it as signs of non-linearity in the data and a pattern means that the model doesn't capture non-linear effects.
- We see no pattern in the plot above. Hence, the assumptions of linearity and independence are satisfied.

ASSUMPTIONS OF LINEAR REGRESSION

3. Normality of error terms





- The residuals more or less follow a straight line except for the tails.
- Let's check the results of the Shapiro-Wilk test.

ASSUMPTIONS OF LINEAR REGRESSION

4. No Heteroscedasticity

Goldfeldquant test

- Ho = Null hypothesis : Residuals are homoscedasticity
- H1 = Alternate hypothesis : Residuals have Heteroscedasticity

shows p> 0.05

Residuals are homoscedasticity

```
[('F statistic', 1.0780199183229537), ('p-value', 0.09280902898661736)]
```

- The p- value is 0.09 > 0.05, hence we can say that the residuals have a constant variance.
- Hence, we can say that all assumptions of our linear regression model are satisfied

FINAL MODEL PERFORMANCE COMPARISON

	LR Skealarn	LR statsmodel
RMSE	0.08348	0.081662
MAE	0.069271	0.070057
R-SQUARED	0.990093	0.989982
ADJ. R-SQARED	0.990066	0.989916
MAPE	1.713701	1.695014

Observations

- The model is able to explain ~99% of the variation in the data, which is very good.
- The train and test RMSE and MAE are low and comparable. So, our model is not suffering from overfitting.
- The MAPE on the test set suggests we can predict within ~1.7% of the used price.
- Hence, we can conclude the model olsmod2 is good for prediction as well as inference purposes.34

FINAL MODEL PERFORMANCE SUMMARY

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

The final predictor variables for the model are following:

All these variables have probability p<0.05 indicating these are all significant variables.

1. 'int_memory': increase in memory increases the used phone price.
2. 'ram': increase in memory increases the used phone price.
3. 'new_price': increase in memory increases the used phone price.
4. 'os_iOS': if phone is having ioS, it increases the used price of phone
5. 'selfie_camera_mp':if phone is having better mp selfie camera, it increases the used price of phone
6. 'days_used': if phone is used for longer time, it decreases the used price of phone
7. 'four_g_yes':if phone is having 4-g technology, it decreases the used price of phone, may be because newer phones have 5-g.

BUSINESS INSIGHTS AND RECOMMENDATIONS

Recommendations based on interpretation of the model input variables

- Company can check used phones with good conditions for
  - new_price <= 600
  - ram>=4,
  - selfie_camera>=8mp
  - days_used is not more than 700,
  - with at least 4-g technology present.
    and then they can sell these phones.
- Most used cell phones are Android devices, they can put them on deal or bundle offers to increase sales.
- ioS is mostly in Apple devices which can be sold as a higher end devices due to its durability and higher new_price.

BUSINESS INSIGHTS AND RECOMMENDATIONS

Comments on additional data sources for model improvement, model implementation in real world, and potential business benefits from model.

- Screen_size, weight variables should accept numbers which are within the correct range of mobile specification.

- It should have a limit check.

greatlearning

Power Ahead

Happy Learning !