ReCell project

Contents

Business Problem Overview and Solution Approach

Background:

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.

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. I have been hired as a Data Scientist by ReCell, a startup aiming to tap the potential in this market. They want me 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.

Data Overview

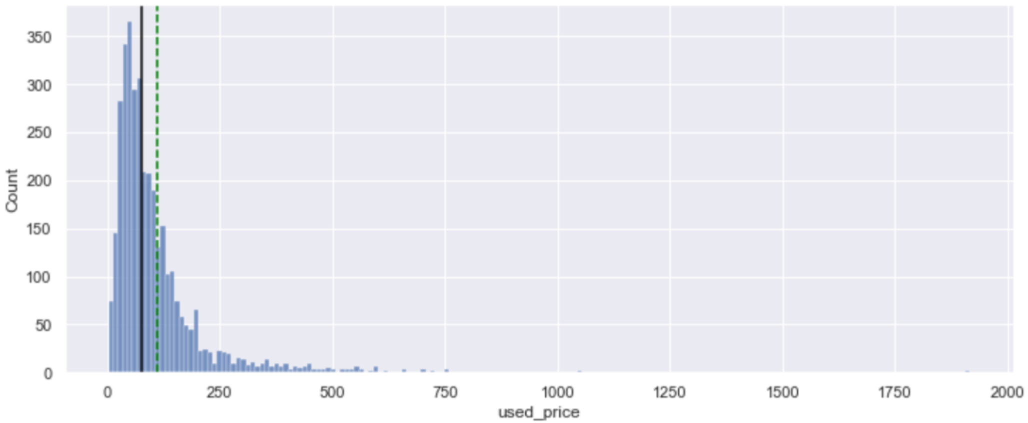
- The data contains the different attributes of used/refurbished phones. The detailed data dictionary is given below
- brand_name: Name of manufacturing brand
- os: OS on which the phone runs
- screen_size: Size of the screen in cm
- 4g: Whether 4G is available or not
- 5g: Whether 5G is available or 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 phone battery in mAh
- weight: Weight of the phone in grams
- release_year: Year when the phone model was released
- days_used: Number of days the used/refurbished phone has been used
- new_price: Price of a new phone of the same model in euros
- used_price: Price of the used/refurbished phone in euros

Data Processing – Initial Steps.

- The dataset has 3571 rows and 15 columns.
- There are null values present for 6 columns : main_camera_mp, selfie_camera_mp, int_memory, ram, battery and weight.
- Among them main_camera_mp has 180 missing values and all others have only 10 or less than 10 missing values.
- The missing values will be replacced in each column with its median.
- There are 34 unique values for brand_name. 4 unique values for os. And 2 unique values 'yes' and 'no' for 4g and 5g.
- The columns brand_name, os, 4g, 5g are object data types.
- These four object columns will be converted to categorical columns.
- All other columns are Numerical with int or float data types.
- The columns 4g and 5g start with numbers. They need to be renamed.
- No duplicate values were found.

EDA

- **Univariate Analysis of used_price**

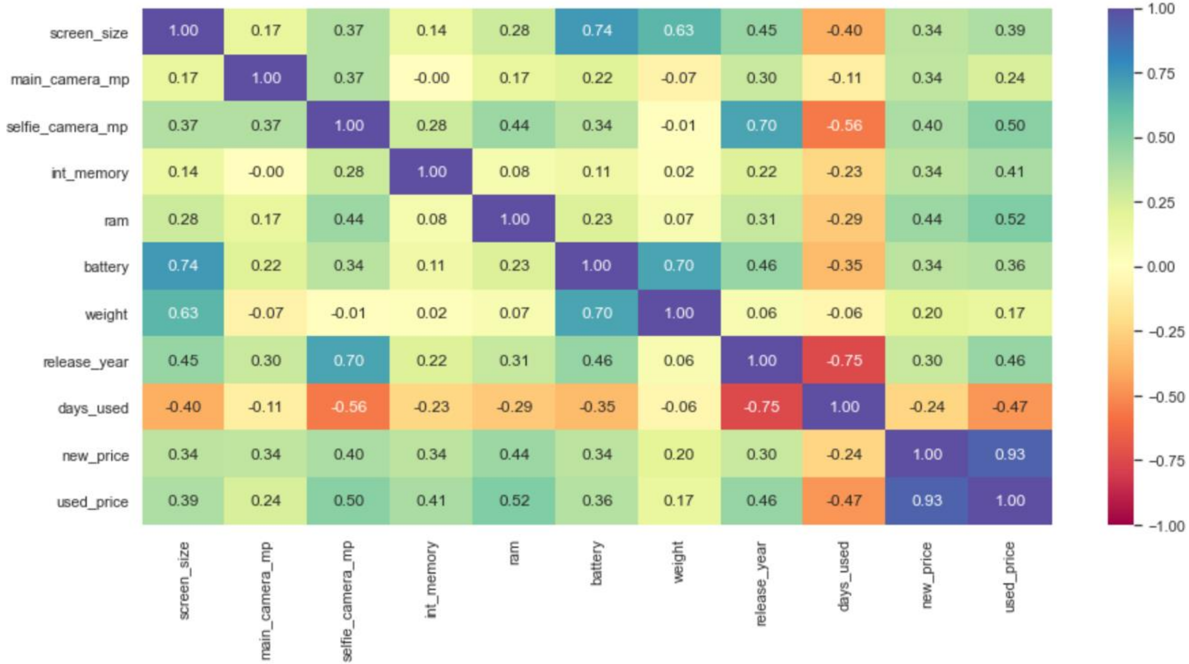


Observations

- The used price of the phone varies from 2.51 euros to 1916.54 euros.
- The mean used price is 109.88 euros and median price is 75.53 euros.
- The mean is higher than the median and the graph is slightly skewed towards the right.
- There are many outliers towards the right indicating many phones with used price more than 250 euros.

Bivariate Analysis

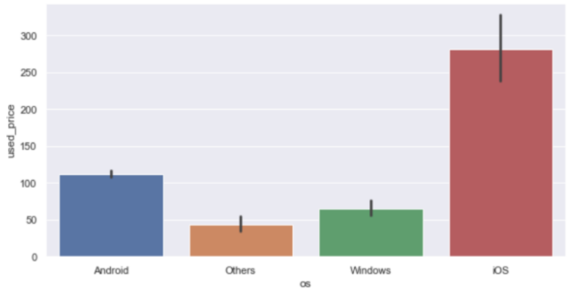
- ## Correlation

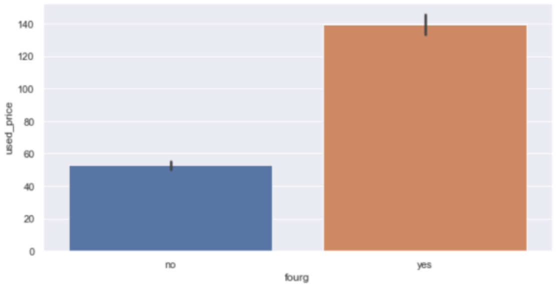


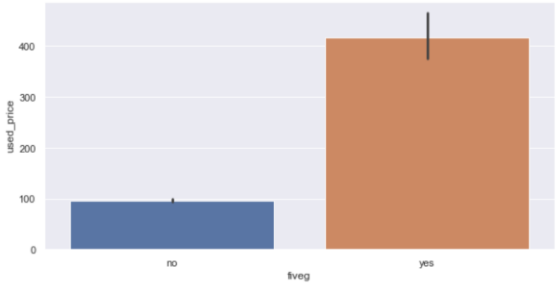
Observations

- used_price has significant positive correlation with new_price and strong positive correlation with selfie_camera_mp and ram. It has weak positive correlation with screen_size, int_memory, battery, release_year.

Relationship of used_price with OS, 4g and 5g







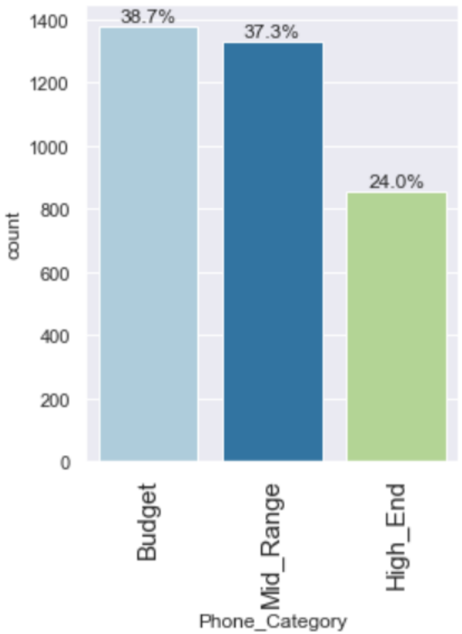
Observations

- Most phones with high used_price are iOS phones.
- Having 4g or 5 g increases the used_price of phones.

Data Processing – Other Steps

- **Column binning**

Used phones are divided into 3 categories – Budget, Mid Range and High End based on their mean used price.

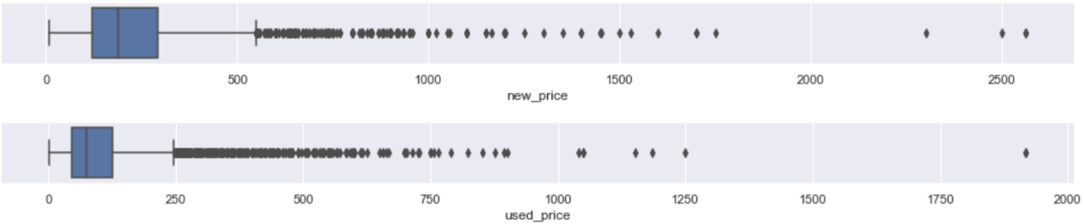


- ## Outlier Detection and Treatment

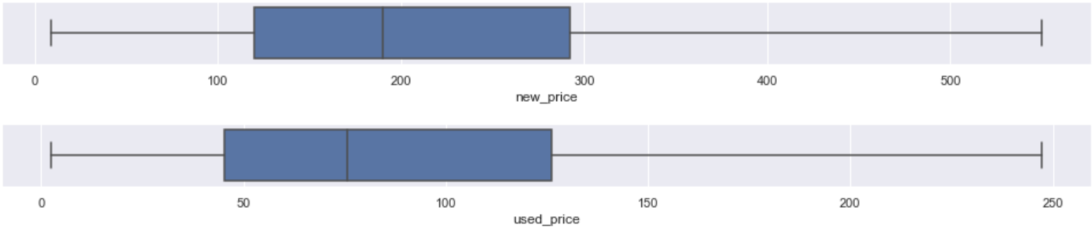
The outliers in the data by flooring and capping.

The below images show the before and after outlier treatment of new_price and used_price.

Before:



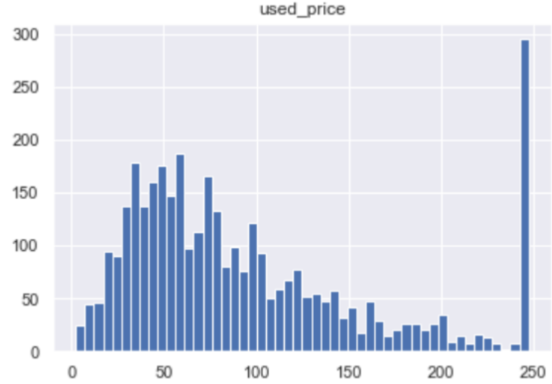
After :



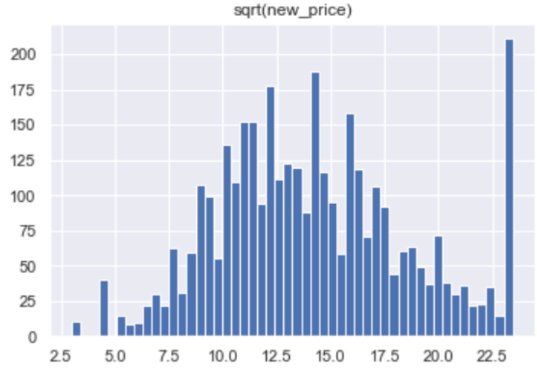
Log transformation

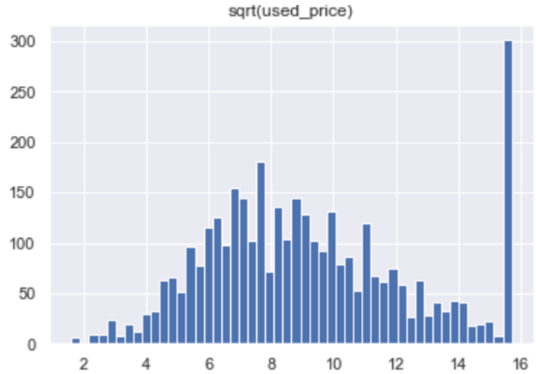
Some features are very skewed and will likely behave better on the log scale. The sqrt function has transformed the new_price and used_price to an almost normal distribution. So we go ahead using the sqrt transformed values.

Before:



After :





Model Performance Summary

Linear Model Building

- First the categorical features are encoded.
- The data is split into train and test to be able to evaluate the model that we build on the train data.
- A Linear Regression model will be built using the train data and we check its performance.
- We use metric functions defined in sklearn for RMSE, MAE, and R2.
- The mean absolute percentage error (MAPE) measures the accuracy of predictions as a percentage, and can be calculated as the average absolute percent error for each predicted value minus actual values divided by actual values. It works best if there are no extreme values in the data and none of the actual values are 0.
- We build the linear regression model using **sklearn** and **statsmodels**.

Checking Linear Regression Assumptions

We checked the following Linear Regression assumptions:

- **No Multicollinearity**

Multicollinearity occurs when predictor variables in a regression model are correlated. This correlation is a problem because predictor variables should be independent.Variance inflation factors measure the inflation in the variances of the regression parameter estimates due to collinearities that exist among the predictors.Since the VIF score for all the variables are less than 5 there is low multicollinearity

- **Linearity and Independence of variables**

Linearity describes a straight-line relationship between two variables, predictor variables must have a linear relation with the dependent variable.The scatter plot of residuals (errors) vs fitted values (predicted values) does not show any pattern. Hence, the assumptions of linearity and independence are satisfied.

- **Normality of error terms**

Error terms, or residuals, should be normally distributed. We check the Q-Q plot of residuals.The residuals more or less follow a straight line except for the heads. As an approximation, we can accept this distribution as close to being normal.

- **No Heteroscedasticity**

If the variance is unequal for the residuals across the regression line, then the data is said to be heteroscedastic. Since p-value < 0.05, we can say that the residuals are Heteroscedastic. Still we will continue with the model as the model is working fine with Budget and Mid Range phones. It could be due to outliers with High End phone price. Also we are provided with limited number of observations that restricts our ability to test.

Model performance evaluation

Let's compare the initial model created with sklearn and the final statsmodels model.

The performance of the two models is close to each other.

	Linear Regression sklearn	Linear Regression statsmodels
RMSE	0.542	0.542
MAE	0.413	0.413
R-squared	0.973	0.973
Adj. R-squared	0.973	0.973
MAPE	4.921	4.921

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.527	0.407	0.974	0.974	4.651

- The model is able to explain ~97% 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 4.65% of the used_price.
- Hence, we can conclude the model *olsmod0* is good for prediction as well as inference purposes.

Final Model Summary

OLS Regression Results

Dep. Variable:	used_price_sqrt	R-squared:	0.973
Model:	OLS	Adj. R-squared:	0.973
Method:	Least Squares	F-statistic:	5533.
Date:	Fri, 01 Oct 2021	Prob (F-statistic):	0.00
Time:	14:35:09	Log-Likelihood:	-2015.0
No. Observations:	2499	AIC:	4064.
Df Residuals:	2482	BIC:	4163.
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	19.3250	20.839	0.927	0.354	-21.538	60.188
screen_size	0.0094	0.005	1.957	0.050	-1.65e-05	0.019
main_camera_mp	0.0009	0.003	0.259	0.795	-0.006	0.008
selfie_camera_mp	0.0252	0.004	6.417	0.000	0.018	0.033
int_memory	0.0018	0.000	4.619	0.000	0.001	0.003
battery	7.005e-06	1.64e-05	0.427	0.669	-2.51e-05	3.92e-05
weight	-0.0002	0.000	-0.329	0.742	-0.001	0.001
release_year	-0.0079	0.010	-0.763	0.445	-0.028	0.012
days_used	-0.0045	6.94e-05	-65.217	0.000	-0.005	-0.004
new_price_sqrt	0.6004	0.004	155.914	0.000	0.593	0.608
os_Others	-0.1594	0.053	-3.026	0.003	-0.263	-0.056
os_Windows	-0.0374	0.081	-0.464	0.643	-0.196	0.121
os_iOS	0.2398	0.098	2.439	0.015	0.047	0.433
fourg_yes	-0.0868	0.034	-2.531	0.011	-0.154	-0.020
fiveg_yes	-0.4078	0.065	-6.265	0.000	-0.536	-0.280
Phone_Category_Mid_Range	0.0331	0.028	1.195	0.232	-0.021	0.087
Phone_Category_High_End	0.0577	0.035	1.670	0.095	-0.010	0.126

Omnibus:	96.524	Durbin-Watson:	1.991
Prob(Omnibus):	0.000	Jarque-Bera (JB):	187.169
Skew:	0.278	Prob(JB):	2.27e-41
Kurtosis:	4.220	Cond. No.	7.34e+06

Business Insights and Recommendations

Conclusions

- new_price has significant relation with used_price. As the new_price increases, the used_price sqrt also increases by 0.60 euros, as is visible in the positive coefficient sign.
- As screen_size, main_camera_mp, selfie_camera_mp and int_memory increases, the used_price increases by not so significant value.
- As the weight, release_year, days_used increases , the used_price decreases as indicated by the negative coefficient.
- The increase in release_year also significantly decreases the used_price sqrt by ~0.25 euros.
- Phones with ios OS significantly increases the used_price sqrt by ~0.23 euros as compared to other OS. For phones with OS listed as "Others" there is significant decrease in used_price sqrt by 0.15 euros as compared to other OS.
- Phones with 4g decreases the used_price by 0.0868 euros as compared to phones without 4g. Phones with 5g decreases the used_price by 0.41 euros as compared to phones without 5g.
- Mid Range Phones and High End phones increases the used_price by 0.033 and 0.057 euros as compared to Budget phones.

Recommendations

- The new price of a phone will heavily determine its used price.
- Phones with iOS os sell with a higher used price than Android or other OS phones.
- Having 4g and 5g in phones are also recommended to get a high used price.
- Selfie camera mp also showed a stron positive correlation with used price.
- Recell Company should bring more of Mid Range and High End phones with higher new price so that they can sell them with higher used price.
- Cameras with good resolution front and back camera should be given more importance.
- Bring in more used iPhones than Android or other os phones.

greatlearning
Power Ahead

Happy Learning !