

USED CAR PRICE PREDICTION

Submitted by:

BALPREET KAUR

**ACKNOWLEDGMENT**

The references I used are:

* Pandas documentation.
* Libraries documentation like:
* [Scikit-learn](https://scikit-learn.org/)
* Seaborn
* Used Stack Overflow website for some technical issues.
* Other tutorial websites for some visibility such as:
  + <https://wellsr.com/python/how-to-make-seaborn-boxplots-in-python/>
  + <https://wellsr.com/python/how-to-make-seaborn-boxplots-in-python/>
  + <https://www.py4u.net/discuss/187401>
  + <https://www.kaggle.com/ammar111/house-price-prediction-an-end-to-end-ml-project>
  + <https://www.researchgate.net/publication/347584803_House_Price_Prediction_using_a_Machine_Learning_Model_A_Survey_of_Literature>
  + <https://www.section.io/engineering-education/house-price-prediction/>
  + <http://www.iraj.in/journal/journal_file/journal_pdf/12-477-153396274234-40.pdf>
  + <https://www.dataquest.io/blog/kaggle-getting-started/>
  + <https://towardsdatascience.com/machine-learning-with-python-regression-complete-tutorial-47268e546cea>
  + <https://www.codegrepper.com/code-examples/python/boxplot+for+all+columns+in+python>
  + <https://ieeexplore.ieee.org/abstract/document/8614000>
* Books consulted:
  + <https://www.researchgate.net/publication/340896273_Machine_Learning_based_Predicting_House_Prices_using_Regression_Techniques>
  + <https://www.researchgate.net/publication/343241145_Housing_Price_Prediction_via_Improved_Machine_Learning_Techniques>
  + <https://www.diva-portal.org/smash/get/diva2:1456610/FULLTEXT01.pdf>

Mu J, Wu F, Zhang A

**Housing Value Forecasting Based on Machine Learning Methods**

Abstract and Applied Analysis, 2014 (2014), pp. 1-7

doi:10.1155/2014/648047.

**INTRODUCTION**

* Business Problem Framing

The business problem is the following one:

The business problem that we need to resolve is to predict the price of a used car.

How this can be related to the real world:

This study is useful whenever there is a study of car prices and can guide the features and parameters that are important to get the best used car by studying its best features.

* Conceptual Background of the Domain Problem

Domain concepts that will be useful for this project is to fully understand the meaning of the features and understand the correlation with the target and study and get full knowledge of the important features concluded by the algorithm. This last acknowledgement should be required after having analysed and conclude the algorithm and get to know the best of the best features concluded by the algorithm to give the best conclusions and recommendations.

* Review of Literature

For review, I checked the attached documentation and identified the names of the features. And researched different websites in order to understand fully the meaning of each of one.

So, basically in the research part, I did try to understand the changes in used car prices over time and how we can describe those changes through its respective used car properties we have in our data frame. And also understand how the online market of used car pricing works as of point of view of Sale and Offers and understand the clients demands regarding the car’s characteristics.

And my main goal for the research was to find, understand and have the better knowledge I can have of the used car features and market in order to better understand the relationship between all features and apply the intensive understanding and knowledge of these features on my used car price prediction project.

It was helpful to have a quick effective research on the concept and how the market works and how the features impact on the price in order to understand better the context.

* Motivation for the Problem Undertaken: IMPORTANT!!

🡪Being honest, I haven’t download the data including enough features due to that the deadline already occurred as of now. I could not download all the data I needed to include in my dataset. But at the same, I was wondering how the ML will work and perform if we give it the data with fewer features.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem:

I did statistical description and studied correlation of the features of the dataset I had for training purposes.

* + In statistical description I checked a first view of skewness. I checked if we had any type of skewness (at first sight) and if so, what type of skewness we had by checking the median, mean and mode.
  + Also, as important study in the statistical descriptive analysis, we checked and got some first insights on the outliers comparing the maximum of each feature with the 75%.

Then in the correlation matrices and heatmaps we did understand the correlation and the quantification of the impact of each feature on our target price value.

And after that, we did an intensive study on the features understanding their important relations and the impacts each feature has on the other one.

I also did print some VIFs values in order to get ride of the features that were mostly correlated to other features and were impacting more due to high multicollinearity.

And after having studied the important features relationships and fully understand the data we have, we started with the process of modelling. As it was a case of price prediction, we followed the basics steps and instructions for linear relationship studies as price has more or less linear type of relation with the other features.

In summary, we did pre-process the data and evaluate the prediction accuracy of the models. It has multiple stages that are required to get the prediction results.

These steps are:

-processing: both datasets will be checked and pre-processed using the methods which have various ways of handling data such as mode, median or mean for each feature.

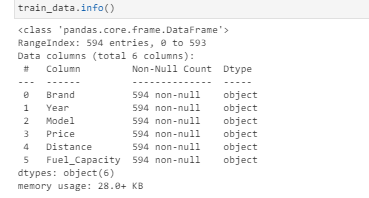
- Data splitting: separating the dataset into two parts n order to train the model with one and use the other in the evaluation. The dataset will be split 75% for training and 25% for testing.

- Evaluation: the accuracy of both datasets will be evaluated by measuring the R2 and RMSE rate when training the model alongside an evaluation of the actual prices on the test dataset with the prices that are being predicted by the model.

- Performance: alongside the evaluation metrics, the required time to train the model will be measured to show the algorithm vary in terms of time.

- Correlation: correlation between the available features and used car price will be evaluated using the Pearson Coefficient Correlation (default correlation method) to identify whether the features have a negative, positive or zero correlation with the used car price.

* Data Sources and their formats



The data was given in the ziP format file and attached to the study case. In the dataset we have a total of 594 entries and 6 columns. Checking the data types, we have a total of 6 features as object type.

So, for our analysis, throughout the process, we reconverted the type of data we have in order to analyse it more in detail. For example, we had to convert the features and encode them in order to add them in the machine learning algorithms.

Several features had object data type when that should have been numerical float or integer data type. So, I needed to convert them to numerical for proper analysis and also for some statistical checks such as outliers.

* Data Preprocessing Done

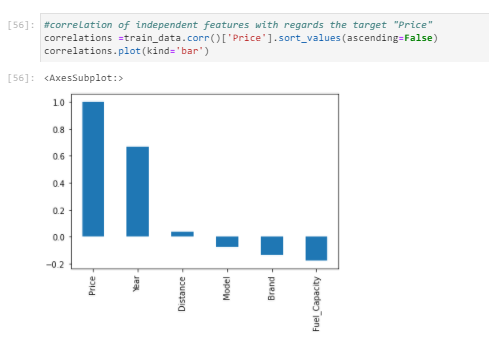
For data pre-processing All we did was the steps that we followed for data preproces and cleaning and basically was to impute the hypen and get rid of pound currency sign and the comma that separates the thousands:



* Data Inputs- Logic- Output Relationships

So, when we study the relation ship between our features with regards our output target, we have several considerations to have in mind:

* + First of all, the data format of the input and output:
    - As we said, we had object as the only format for our input data.
    - The problem was basically with object type data when considering them for the relationship purposes. And for these, we had to label encode them in order to convert them in the right format for the correlation studies.
  + Relation and affection between the input and output:
    - When checking the relationships between the features and the target, we see some feature like Year, Fuel\_capacity, Brand that are more correlated and have more possibility to impact more on our target variable used car Price compared to others.



* Hardware and Software Requirements and Tools Used

I used Jupyter-notebook and Python 3 language.

Imported required libraries such as:

- pandas

- numpy

- matplotlib.pyplot

- seaborn

- sklearn

For modelling, we imported different algorithms, pre-processing, splitting tools, data transformation and metrics from different libraries:

* From sklearn.linear\_model imported LinearRegression
* From sklearn.linear\_model import Lasso (Lasso is a type of regression that uses shrinkage)
* From sklearn.tree imported DecisionTreeRegressor
* From sklearn.neighbors imported KNeighborsRegressor
* From sklearn.svm imported SVR
* From sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor, AdaBoostRegressor

All these previous imports were executed in order to fit and train the data.

* From sklearn.model\_selection imported train\_test\_split

The train\_test\_split was used to split the data in the training and testing set.

* From sklearn.preprocessing import LabelEncoder

LabelEncoder helped me to convert the object type data to numerical for statistic analysis including correlations, plotting purposes and machine learning algorithm training.

* From scipy.stats import skew 🡪 for skewness calculation.
* From statsmodels.stats.outliers\_influence import variance\_inflation\_factor 🡪 to calculate VIF and check multicollinearity.
* From scipy.stats import zscore 🡪 to calculate zscore and check the outliers.
* From sklearn.preprocessing import power\_transform🡪 to transform and reduce the skewness.
* From sklearn.preprocessing import StandardScaler🡪 to normalize the data and put it in the same scale for better comparison.
* From sklearn.metrics import r2\_score, mean\_absolute\_error,mean\_squared\_error🡪 metrics error to calculate for algorithm comparison and will help us to select the best algorithm for our case.
* From sklearn.model\_selection import GridSearchCV
* From sklearn.model\_selection import RandomizedSearchCV

The gridsearch CV and randomizedsearch CV is used to get the best CV we can get in order to get the best scores through our best algorithm.

* From sklearn.model\_selection import cross\_val\_score

Cross\_val\_score helps us to calculate the score of the cross validation.

* Executed %matplotlib inline to call the predefined functions that will plot and default style the plot for you.

Also did import warnings in order to ignore them:

import warnings

warnings.filterwarnings('ignore')

And also imported pickle for the model saving.

**Model/s Development and Evaluation**

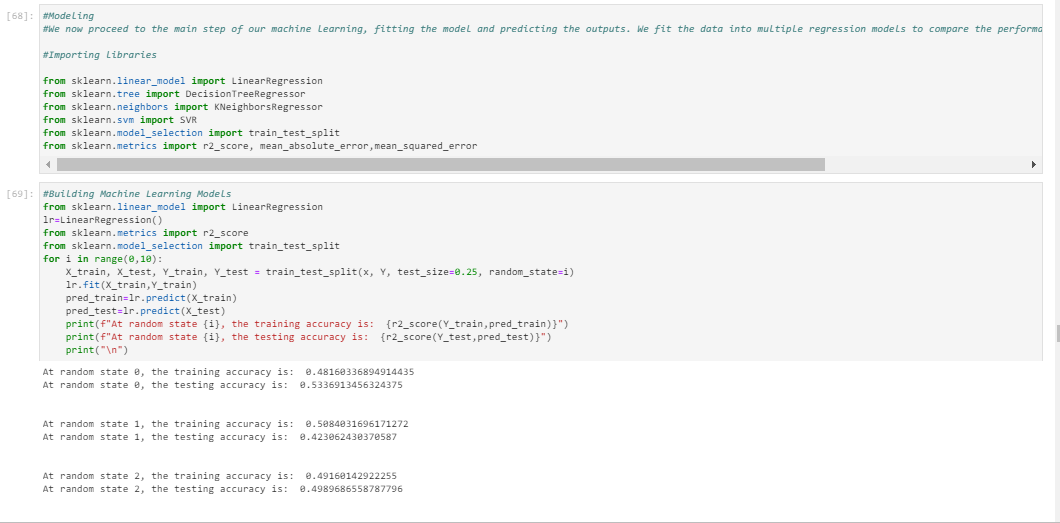
* Identification of possible problem-solving approaches (methods)

The main approach I decided to follow is to use linear algorithm

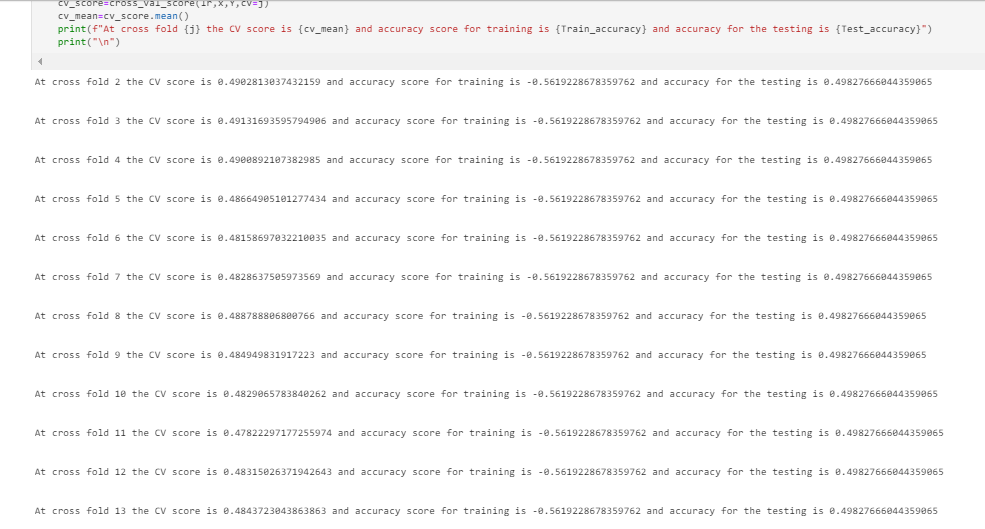
and a few other regressions algorithms as the model should do a prediction of a continuous target variable.

* Run and Evaluate selected models

Then, after checking and doing some statistical descriptive analysis, we decided starting with simple Linear regression modelling.



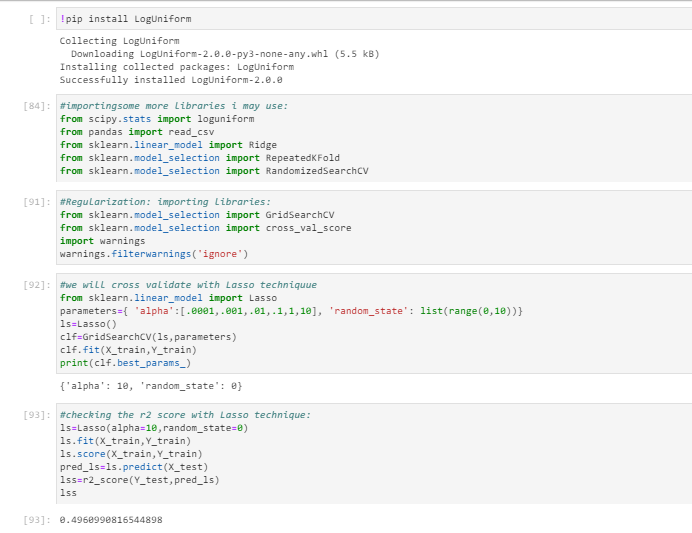






As you see above, we trained the algorithm, we did some study on the best CVs and printed the results score and checked for MAE, MSE and RMSE so that we can decide which algorithm is working best for our study case.

After having printed the results score and errors metrics for linear regression, we tried other algorithms in order to check if we can get better results than the linear regression. We did some Lasso Regression too in order to improve the current results and scores.



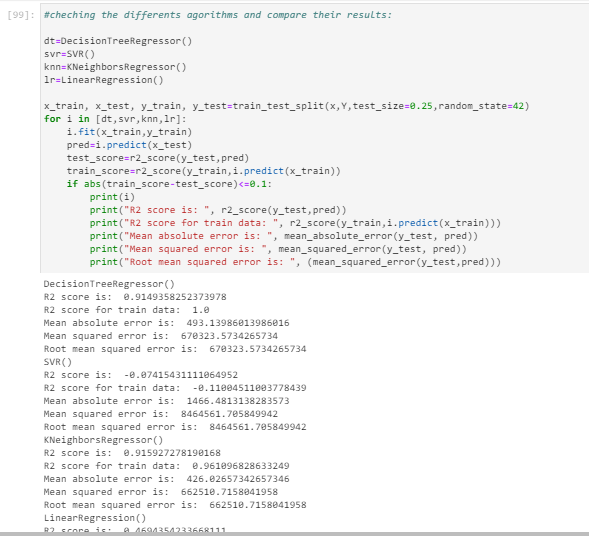
I also tried some ensemble techniques such as Random Forest Regressor and compared the results:



I also decided to check and compare the results of:

* + DecisionTreeRegressor
  + SVR
  + KNeighborsRegressor
  + LinearRegression
  + RandomForestRegressor
  + AdaBoostRegressor
  + GradientBoostingRegressor

Checked these algorithms metrics and compared the results and scores:



Also compared the results of RFR, ABR and GBR:

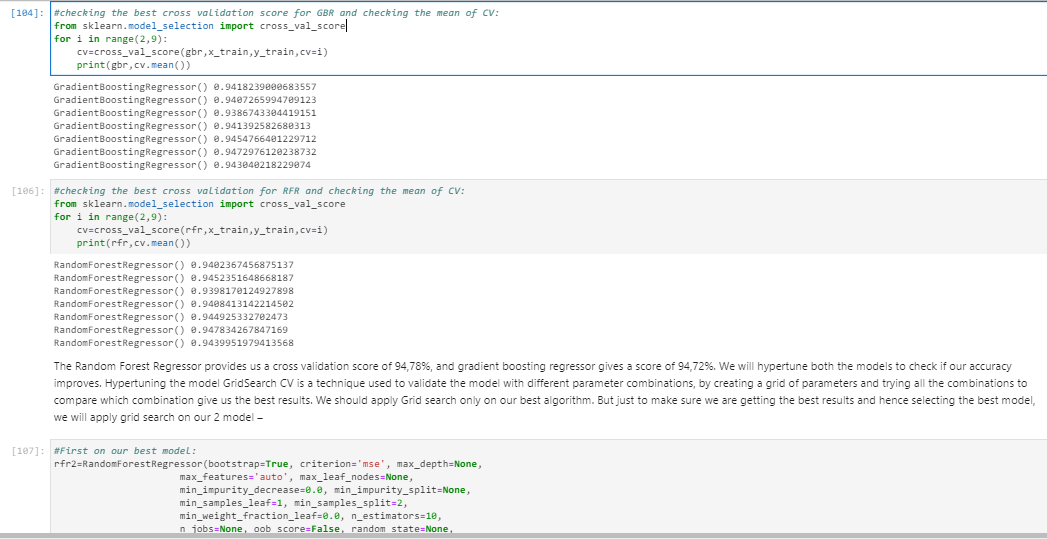


As we see GBR and RFR are the best algorithms that we can get from the previous steps, I decide to check the r2 score and errors metrics MAE, MSE and RMSE on these two and compare the results of both:

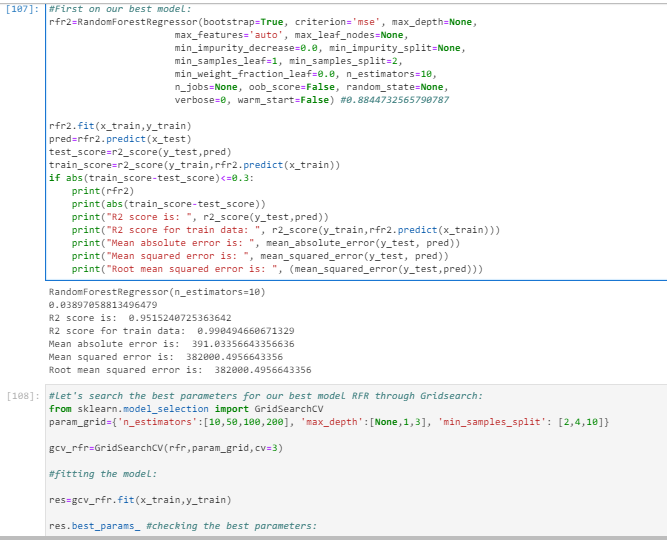


GBR model gives us the best accuracy till now, with an R2 score of 95,96 which is better than K-Neighbors and the model is not overfitting as well. Mean Absolute error for this model is ~368 and RMSE ~ 317k.

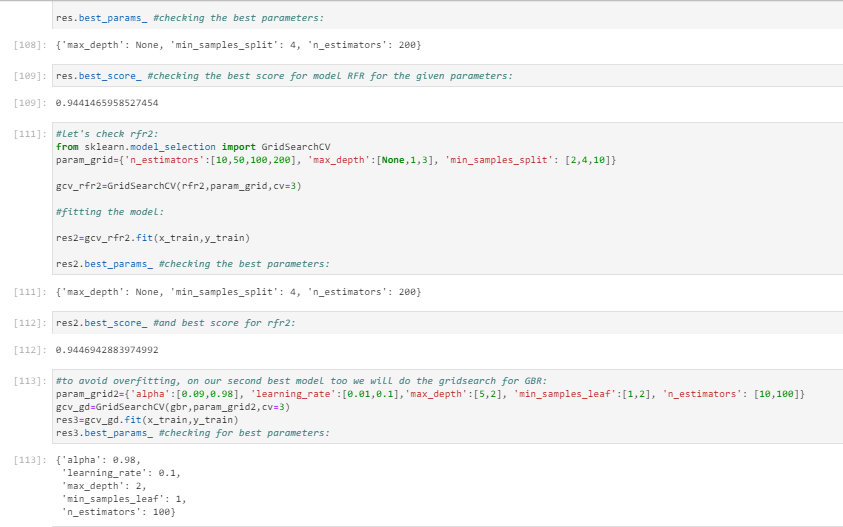
Cross Validation We will perform the cross validation of our 2 model to check if the models have any overfitting issue by using k-folds. We test the cross validation for Random forest and Gradient Boosting Regressor.



The Random Forest Regressor provides us a cross validation score of 94,78%, and gradient boosting regressor gives a score of 94,72%. We will hypertune both the models to check if our accuracy improves. Hypertuning the model GridSearch CV is a technique used to validate the model with different parameter combinations, by creating a grid of parameters and trying all the combinations to compare which combination give us the best results. We should apply Grid search only on our best algorithm. But just to make sure we are getting the best results and hence selecting the best model, we will apply grid search on our 2 model.

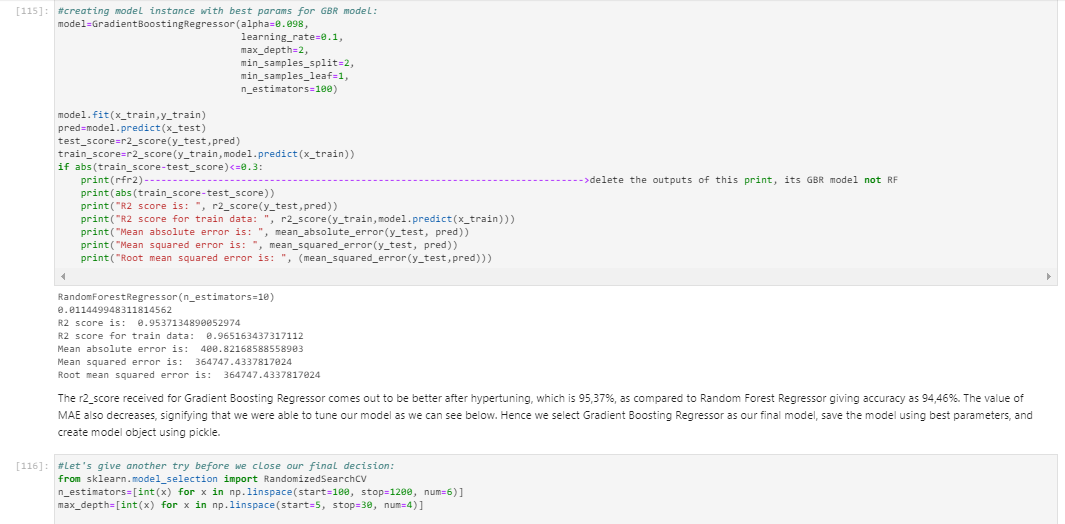


Here above screenshot we clearly see that our algorithm RFR has better score results (95.15%) to offer than the algorithm GBR which has score of 94.72%.



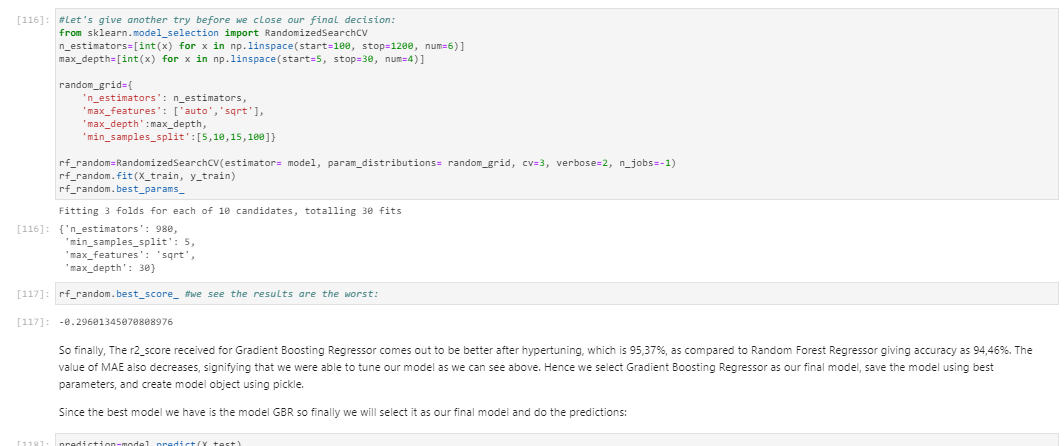


Let’s now create the best instance of the model GBR and check the score results and the metrics of error: MAE, MSE and RMSE alongside with R2 Score:



The r2\_score received for Gradient Boosting Regressor comes out to be better after hypertuning, which is 95,37%, as compared to Random Forest Regressor giving accuracy as 94,46%. The value of MAE also decreases, signifying that we were able to tune our model as we can see below. Hence we select Gradient Boosting Regressor as our final model, save the model using best parameters, and create model object using pickle.

But trying again, we get that:

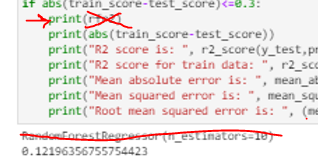


Trying to do a Randomized Search CV on the same best model till now GBR.

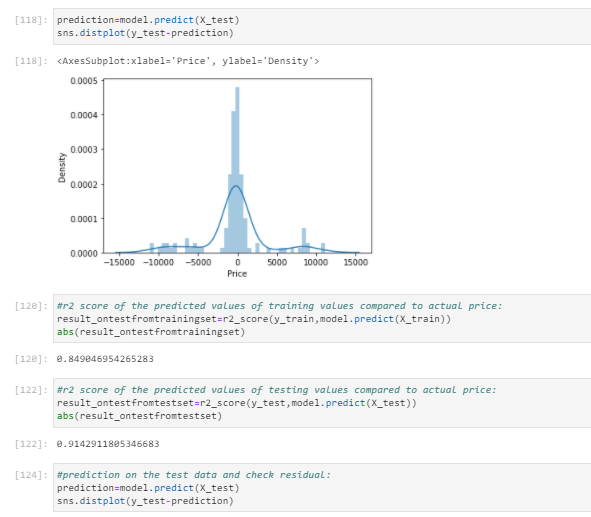
So finally, The r2\_score received for Gradient Boosting Regressor comes out to be better after hypertuning, which is 95,37%, as compared to Random Forest Regressor giving accuracy as 94,46%. The value of MAE also decreases, signifying that we were able to tune our model as we can see above. Hence we select Gradient Boosting Regressor as our final model, save the model using best parameters, and create model object using pickle.

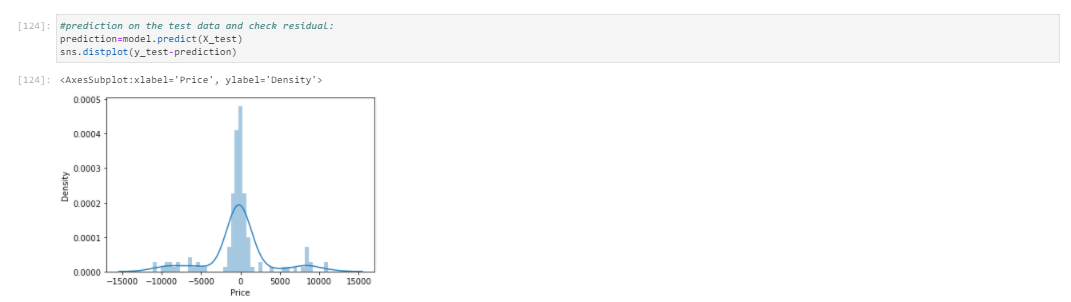
Since the best model we have is the model GBR so finally we will select it as our final model and do the predictions.

I just realized the error in the last moment when writing this report and the error I did is to print the name variable rfr2 when I should have put gbr there:

(I should have printed the name of GBR instead of printing the RFR, [😅](https://emojipedia.org/grinning-face-with-sweat/))

Ok, so as we know our best algorithm is GBR and we did the Randomized Search CV on the same best model which is till now GBR.





Even after checking both algorithm scores, we clearly see that GBR scores are way better. We have achieved an r2\_score value of 91% on the testing set, meaning that we are actually able to predict values quite near to the actual prices, for majority of the rows. A glimpse of our resulting dataframe is attached below for the testing set. We load the test file, apply all the data modeling processes and operations on our test data similar to what we did with the train data, and then make the final prediction using the saved model object as follows:



* Key Metrics for success in solving problem under consideration

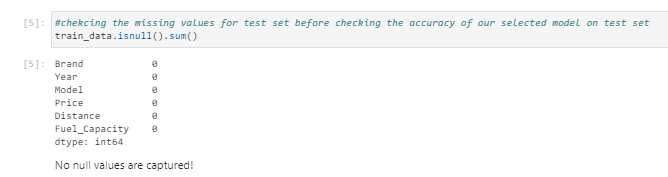
Basically, I used the 3 errors metrics and R2 score alongside the test score in order to compare and choose the best algorithms as shown below:

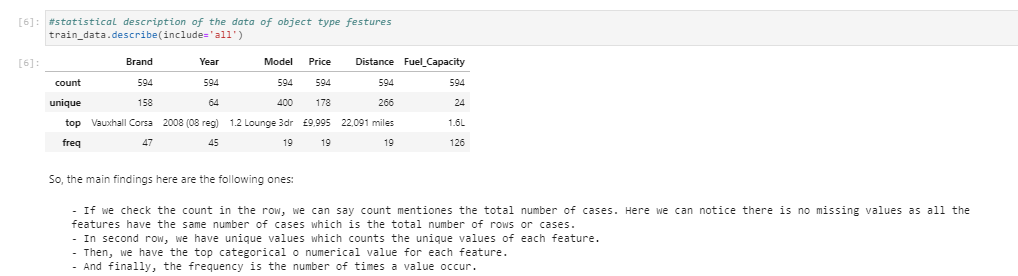


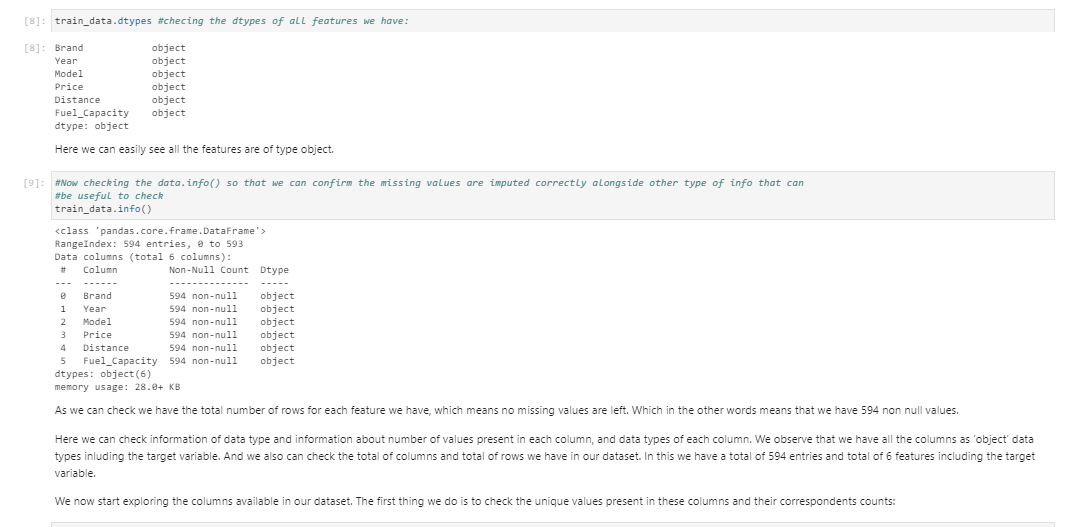
I considered all the results and scores one by one and compared this way all algorithms and chose the best one which is GBR.

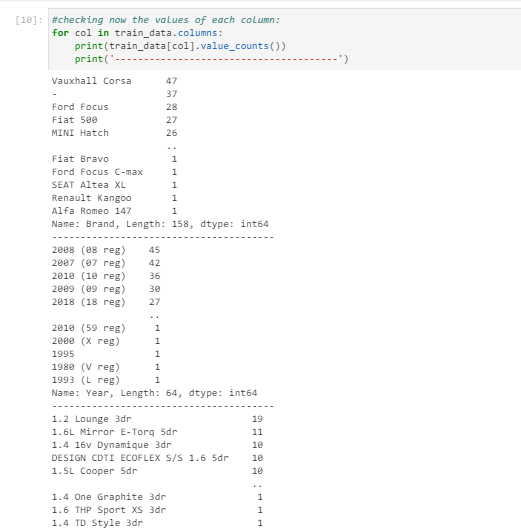
You can easily understand the all the process I did when you read my python script as it has complete similar comments on most of the main steps and everything written here will make more sense and you will understand all the steps and considerations commented here. Please go through all the comments in the given code script and will also make it easy to understand this report.

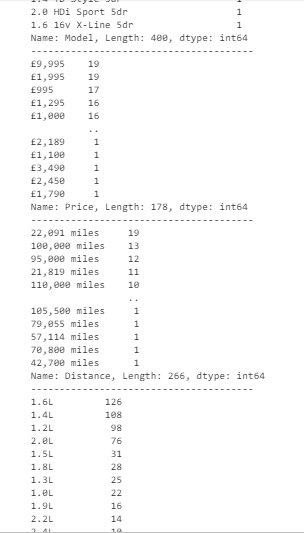
Let’s check and revise the main findings:

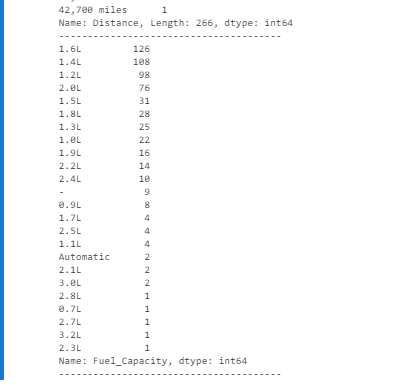












Let's highlight the most visible characteristics of this data:

So, first of all, we see Brand, from where we can highlight that Vauxall Corsa, Ford Focus, Fiat 500 and Mini Hatch are the 4 brands with highest number of used cards offered in the website and included in this dataset.

Secondly, regarding the Year, there are no specific dates when the used cars are not sold; the distribution is almost similar for all dates, at least for the first 5 values. But we have Years which are not that much important but we can highlight that the Years with highest cases of used cars are 2008,2007,2010,2009,2018.

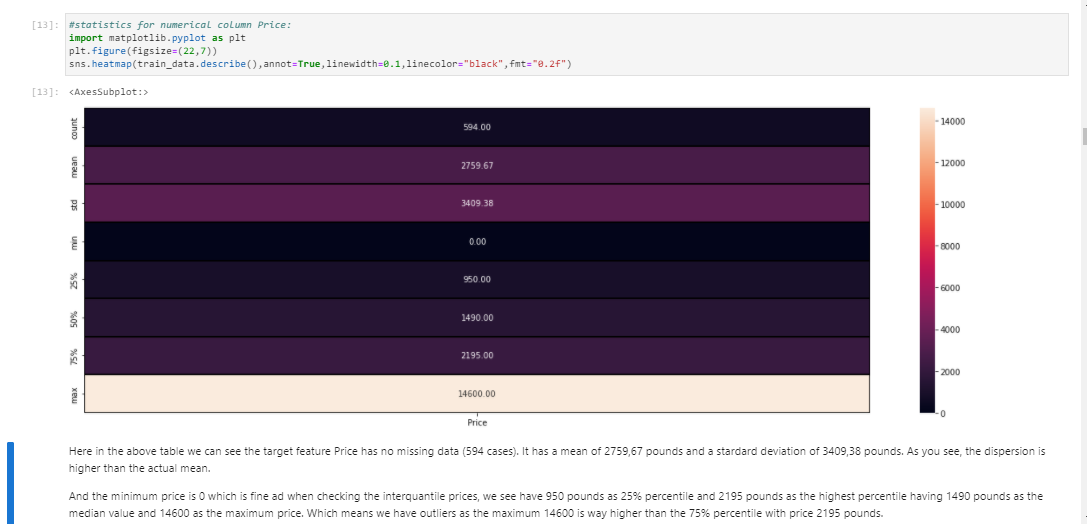
Regarding the Model, we can say 1.2 Lounge 3dr, 1.6 Mirror E-Torq 5dr, 1.4 16v Dynamique 3dr the 3 models with highest number of used cars when we are studying the model of the used car. And 1.4 One Graphite 3dr, 1.6 the sport XS 3dr,1.4 TD Style 3dr,2.0 HDi Sport 5dr and 1.6 16v XLine 5dr have the minimum count of used cars in their website. And when checking the 9995, 1995 995, 1295 and 1000 as the most offered price for a used car and which means the maximums price.

And regarding the Distance, we have 22091 miles and 100000 miles as the 2 distances which are mostly used in our dataset. If you check more in detail, the first route is almost 46% higher when comparing with the second one.

When comapring the Fuel capacity , we have a total of 126 cases if we check all the cases with only 1.6L of capacity of fuel of a used car. And we have 108 cases of cases where we have a capacity of 1.4L of fuel. Interesting!. Which means people in this dataset had more used cars with 1.6L of fuel capacity when compared to other options. And when comparing the third option, we see we have 98 cases which is only 16% when comparing with others. In summary, majority of the used cars have either fuel capacity of 1.6L or 1.4L.

* Visualizations

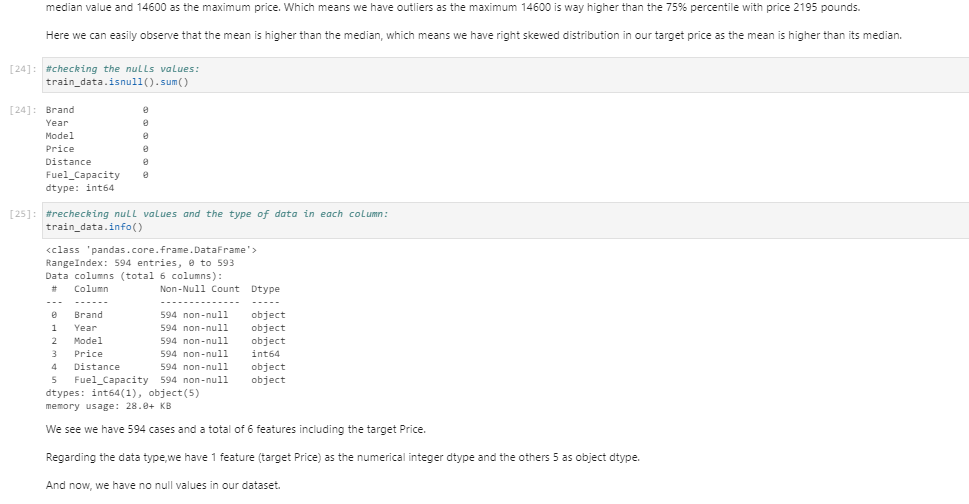
First of all we plotted descriptive details of our data:



Here in the above table we can see the target feature Price has no missing data (594 cases). It has a mean of 2759,67 pounds and a stardard deviation of 3409,38 pounds. As you see, the dispersion is higher than the actual mean.

And the minimum price is 0 which is fine ad when checking the interquantile prices, we see have 950 pounds as 25% percentile and 2195 pounds as the highest percentile having 1490 pounds as the median value and 14600 as the maximum price. Which means we have outliers as the maximum 14600 is way higher than the 75% percentile with price 2195 pounds.

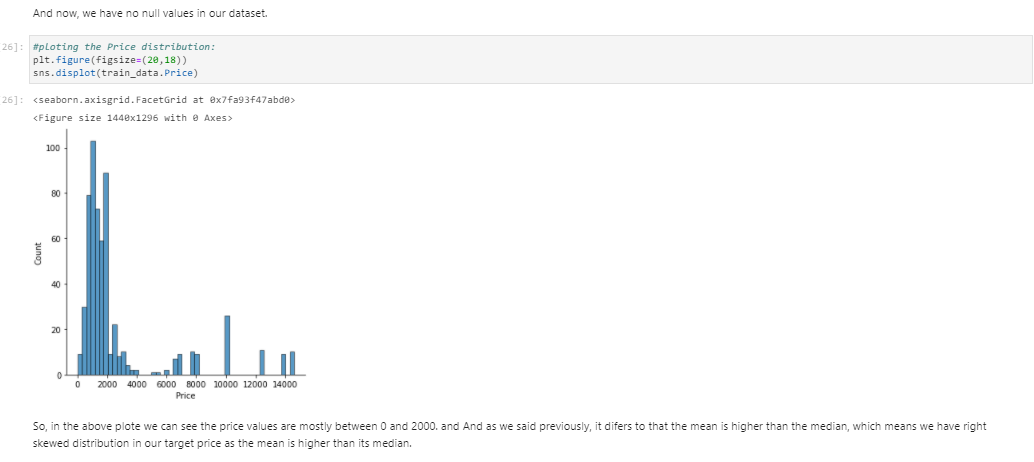
Here we can easily observe that the mean is higher than the median, which means we have right skewed distribution in our target price as the mean is higher than its median.



We see we have 594 cases and a total of 6 features including the target Price.

Regarding the data type,we have 1 feature (target Price) as the numerical integer dtype and the others 5 as object dtype.

And now, we have no null values in our dataset.



So, in the above plote we can see the price values are mostly between 0 and 2000.

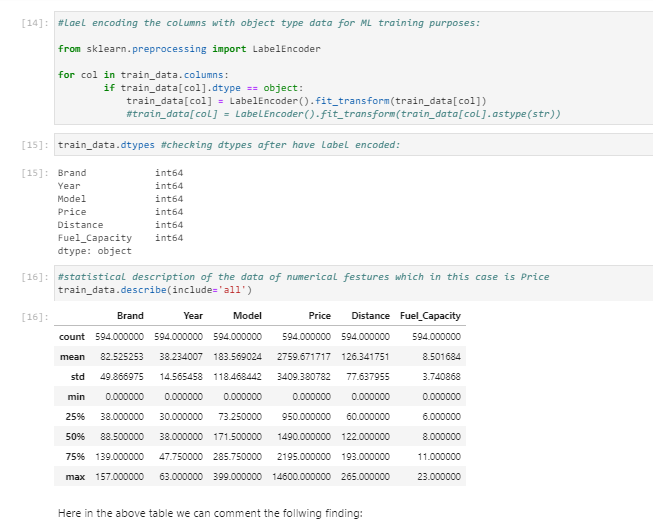
And as we said previously, it difers to that the mean is higher than the median, which means we have right skewed distribution in our target price as the mean is higher than its median.



Same as previous plot:

So, in the above plot we can see the price values are mostly between 0 and 2000.

And as we said previously, it differs to that the mean is higher than the median, which means we have right skewed distribution in our target price as the mean is higher than its median.



Here in the above table we can comment the follwing finding:

Here in the above table we can see all the features have no missing data (594 cases).

As you see, the dispersion of price is higher than the actual mean.The remaining features have mean higher thn the dispersion.

And the minimum of all features is 0 which is fine

ad when comparing the 75% quantile and the maximum values, we see all the features including our target haver its maximum higher than its 75% quantile, Which means we have outliers as the maximum is way higher than the 75% percentile.

Here we can easily observe that the mean of Price is higher than the median, which means we have right skewed distribution in our target price as the mean is higher than its median. It happens the same with feature Year, Model, Distance and Fuel\_Capacity.Except the feature Brand which has the median higher than mean, which means for feature Brand we have a left skewed distribution.



Let's comment and highlight the relations where we can from the above scatterplots:

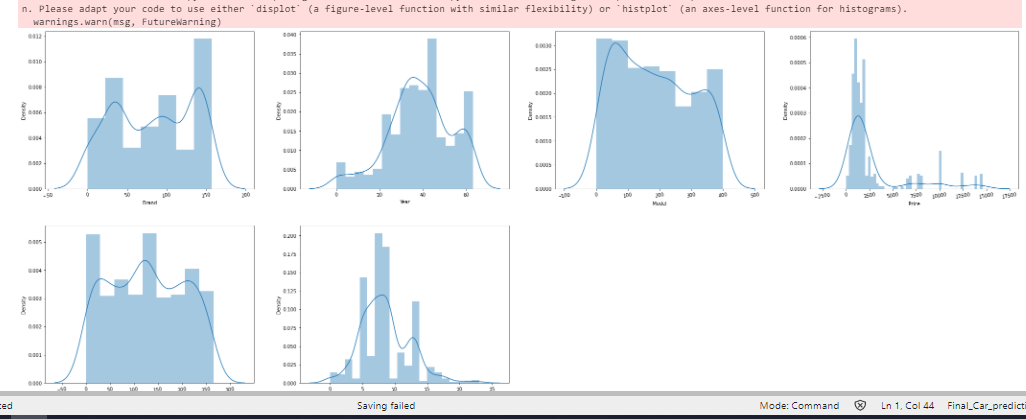
Except the Price itself, all other features have no linear pattern in the above plots.

All the features have outliers pattern in the above plots.

And if we check the Fuel Capacity in detail, we see that the less Fuel Capacity is, the high is the dispersion of the data values.

Let's distplot and check if we have the same conclusions!





First of all, we can hihlight the following observation of all the distribution:

- We have distribution near to Normal distribution except a few features.

- For feature Brands, as we said, we have Vauxhall Corsa with the highest number of cases.

- For feature Year, as we said, we have 2008 and 2007 have the highest number of cases.

- For feature model, almost all the models have a quite similar number cases, close to each other.

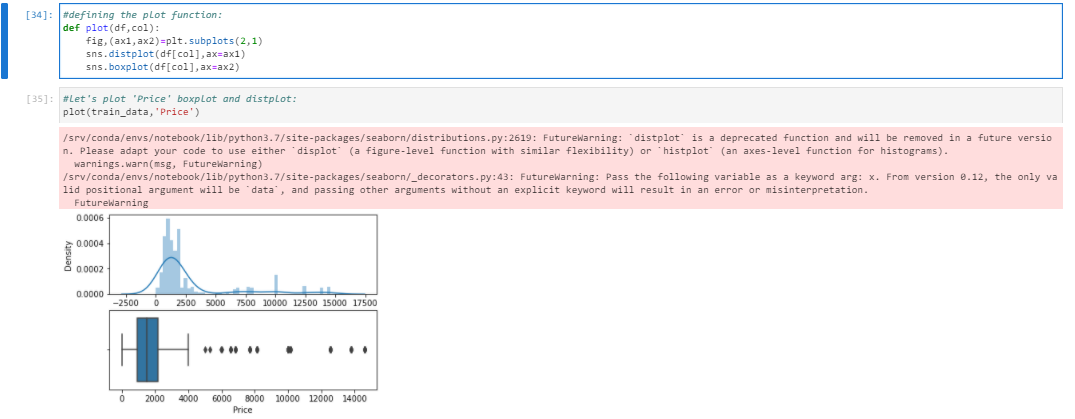
- Price has a right skewed distribution as its mean is higher than its median.

- For feature Distance, we see the 22,091 miles,100,000 miles,95,000 miles,21,819 miles are the 4 with highest number of cases.

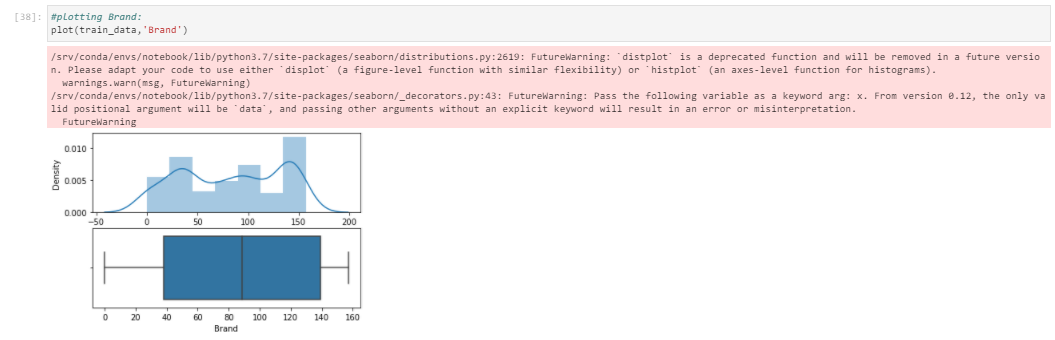
- For feature Fuel capacity, we see we have a distribution that is centered throughout the different Fuel Capacity having median, mode and mean near to value between 1.4L and 1.6L.



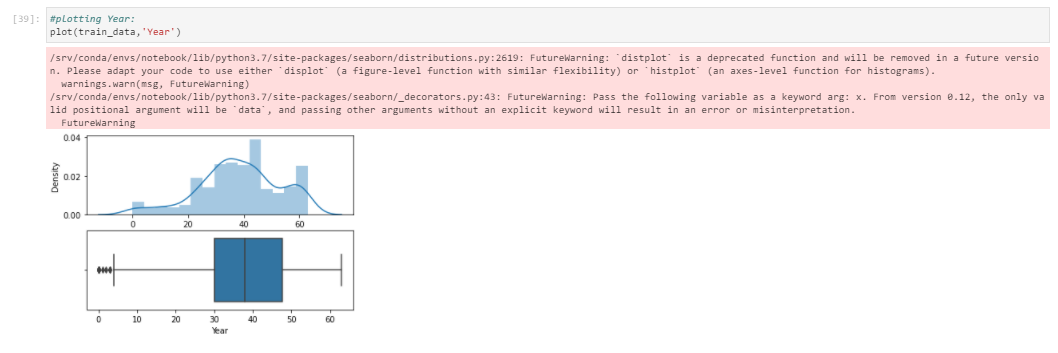
Regarding the price target, the majority of the USED Cars have price range between around 0-2.5k and then the cars number decreses after the price range over 2.5k. Used cars having prices greater than 15k are quite less. Price range is skewed towards right. As we said earlier in the analisis, we have right skewness in the target price. We will now compare several independent features with ‘Price’ column to check their impact:



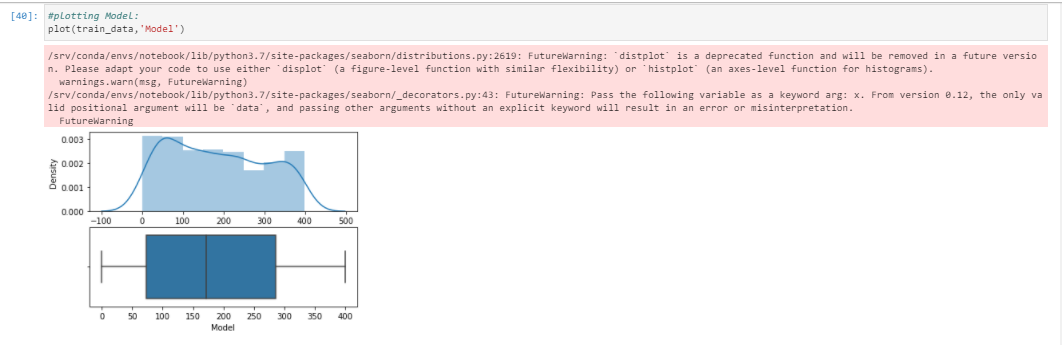
If we check density, as we said, we have range of 0 and 2.5k. So, the distribution has right skewness. And then, if we check boxplot, we see a median on around 1.6k which means outliers taking place between 4k and 17.5k are impacting way more than the other cases.



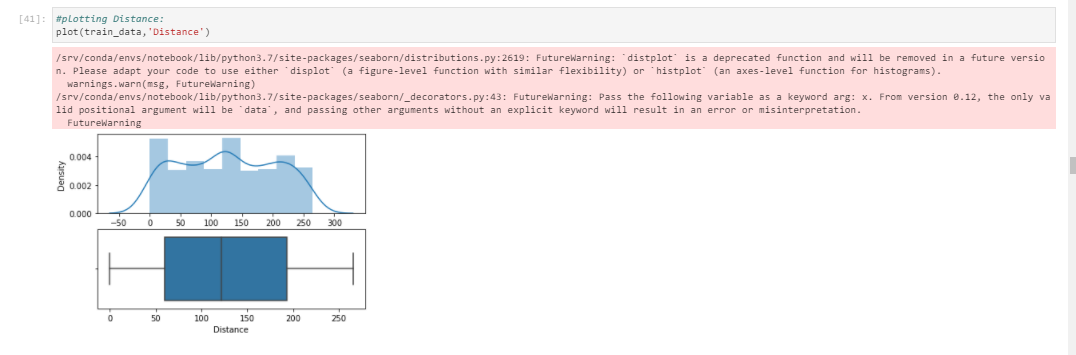
So, the distribution is close to normal distribution. Checking the median doesn't make any much sense as this feature should be studied as categorical data.



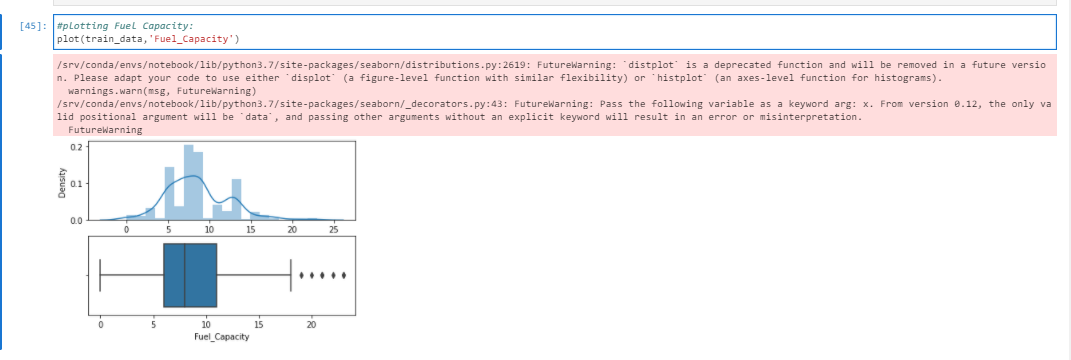
So, the distribution has a bit of left skewed distribution but it isnt meaningful as it should be treated as categorical data or date-serial type data. Checking the median doesn't make any much sense as this feature should be studied as categorical data.



So, the distribution has a bit normal distribution. Checking the median doesn't make any much sense as this feature should be studied as categorical data.

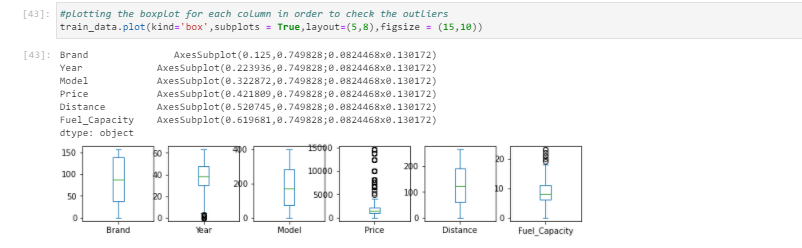


If we check density, as we said, we have the feature Distance with a range of 0 and 250k. So, the distribution is more likely to be a semi normal distribution. And then, if we check boxplot, we see a median on around 124k and there is no outliers as seen in the boxplot.

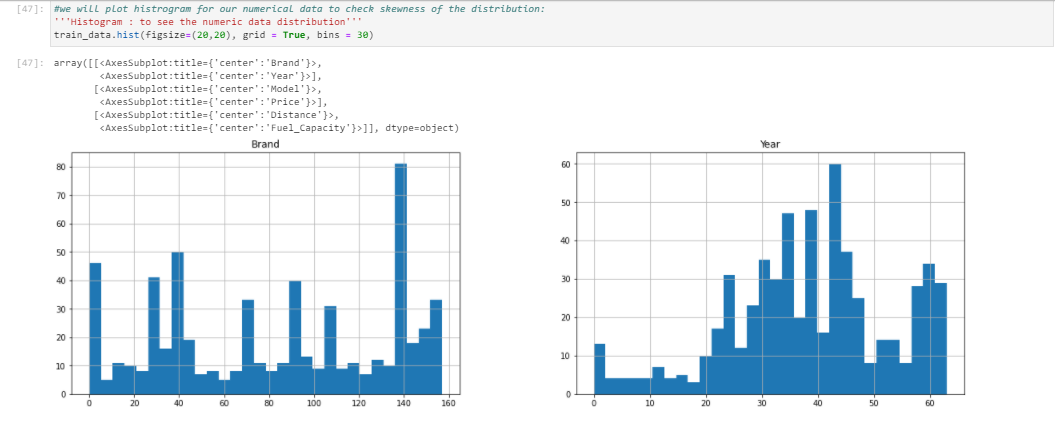


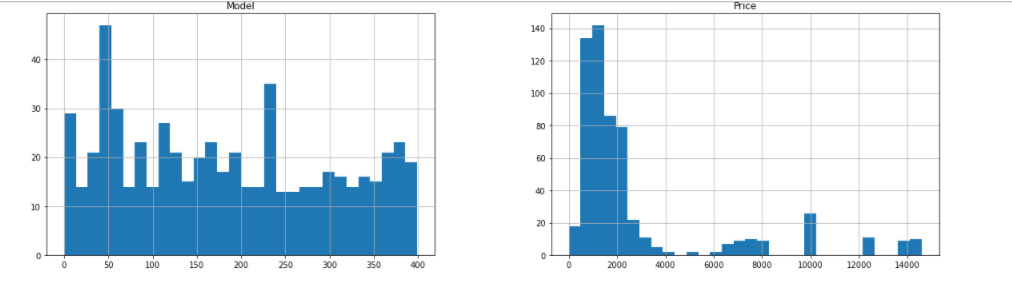
As we said earlier, Fuel Capacity has a normal distribution as its median is very close to the mea.

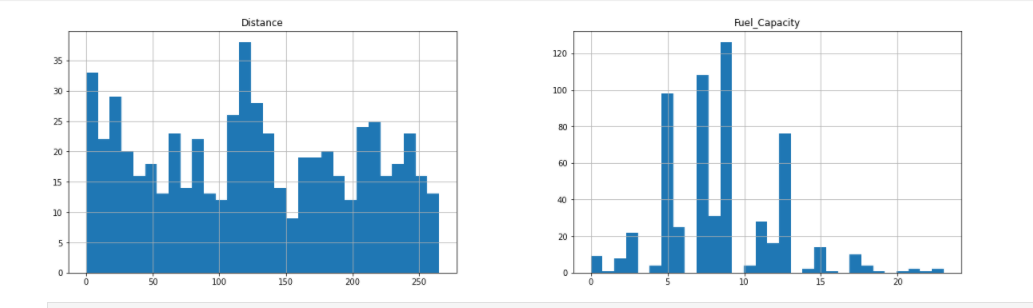
And in boxplot, we have outliers when label encoded code is higher than 18 which we should have not label encoded, instead changed to data type numerical float instead of label encoding it.



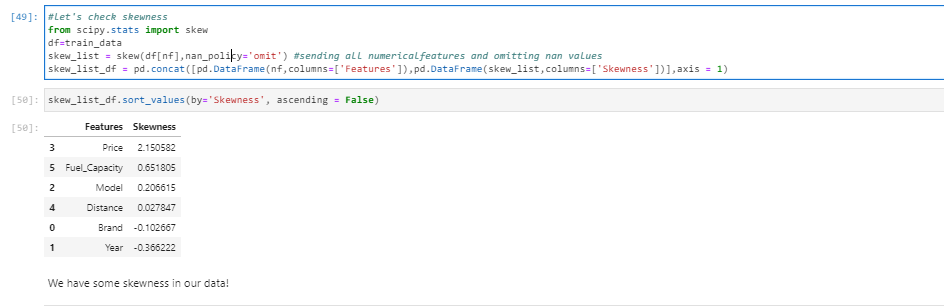
So, in the boxplot plotted above, we can easily see we have outliers in almost all our features except Brand, Model and Distance.



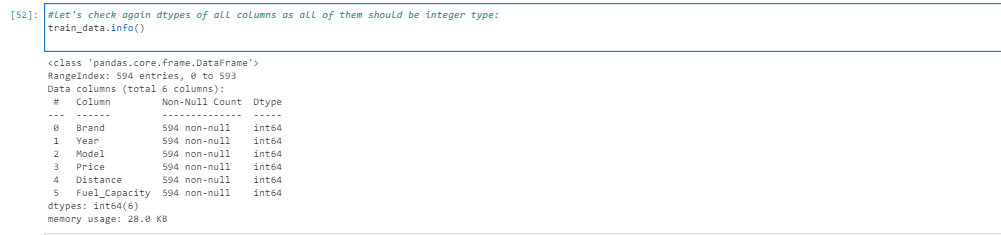




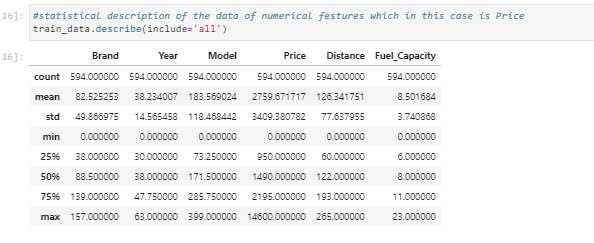
#we see skewness in the above grafs. So, let's go and select numerical columns and check their corresponding skewness:



We have some skewness in our data!



We can see that, all 6 columns are of numeric data types. Out of 6 variables, 5 are predictor variables and the one 'Price' is the target variable. The dataset is fine since there are no null values and all data is in type numerical integer type. So, let's move forward!

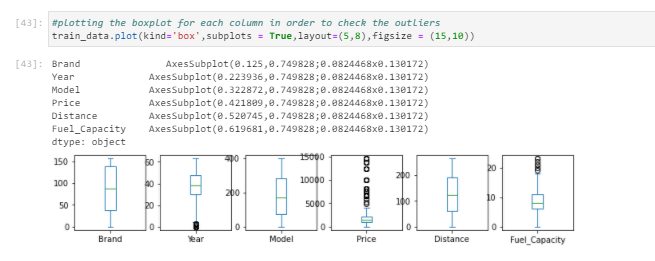


From the above plots we could check at first sight if we had any skewness or outliers. Also checking the total number of instances, we could check and confirm that we do not have any missing values. With the above plot we also can check if we have same or different scaling of the features. If so, apply any kind of normalization technique such as standard scaler or tmin-tmax scaler.

On the one hand, for all the features where the mean is smaller than the median, we can say it has left skewed data which also means negative skewness. On the other hand, if a feature has the mean higher than the median, than it is right skewed which is also called as positive skewness.

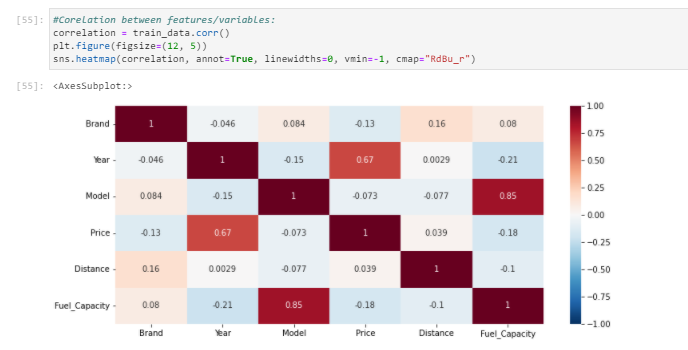
For outlier check at first sight, we can notice it when we have larger values for the maximum compared to 75% percentile values.

We also plotted box plots for all our features including the target in order to check outliers in all the features:



The circles are the outliers for each feature. Almost all the features have outliers including our target variable which is one of the variables that has high number of outliers. Only Brand, Model and Distance do not seem to have outliers.

Then, we also plotted the correlation between features through a heatmap in order to see easily the best correlations between all our features including our target “Price” of used cars:



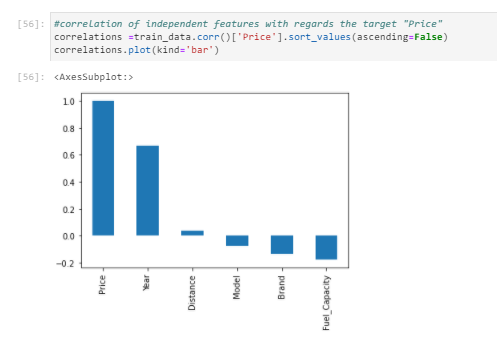
If we take out the correlation equal to 1, we see the highest correlation are:

- Feature Model is highly correlated with Fuel Capacity.

- And regarding the target Price, we see the feature that are highly correlated with it is the Year.

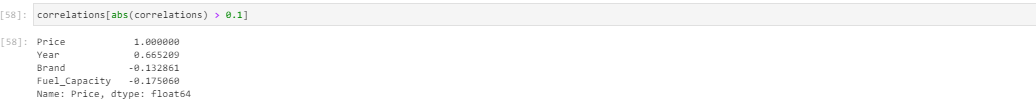
Let's check plot of the correlation of the features with regards the target feature in the following code:

If we concentrate on the target and check the best correlations against it, then we will see that the following features are considered the best correlated features within our dataset:



As we can see, as we said, Year has the most impact on our Price target feature. And then, we have other features like Fuel Capacity and Brand which impact more on price in second position and third position respectively.

Now let's check the features that has most impact on price:



As we said, we have Year as the main independent feature that impacts the most on the Price of the used cars for the dataset we have. ANd then, Fuel Capacity and Brand have impact on price in second position and third position respectively.

So here are the answers to the following questions:

• Which variables are important to predict the price of variable?

When checking the relationships between the features and the target, we see some feature like Year, Fuel\_capacity, Brand that are more correlated and have more possibility to impact more on our target variable used car Price compared to all the available features in our dataset.

• How do these variables describe the price of the used car?

These variables have higher impact on the target then the others when comparing the percentage of correlation. So, the higher correlation, more probability to have higher impact on the Price of a used car.

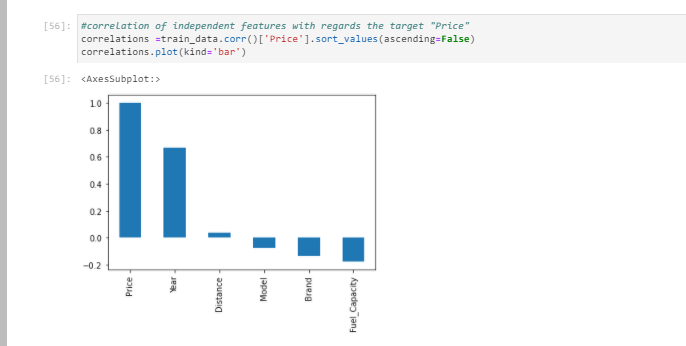
* Interpretation of the Results

First of all, we took off hypen, pound sign and comma from the Price feature.

Then, we had skewness, multicollinearity and outliers that we had in all the columns. We looked and managed these 3 statistical measures through some power transformation yeo-johnson method, dropping irrelevant columns through VIF calculation and calculating ZScore for the 3 statistical measures respectively.

Then, after analysing several different visualizations, pre-processing and modelling, we got the following results:

In machine learning, it is a general practice to rely on a correlation matrix to decide what features to be incorporated into our models. We noticed we had several features such as Year, Fuel\_capacity, Brand which are the features that have higher impact comparing all the available features.



These variables have higher impact on the target then the others when comparing the percentage of correlation. So, the higher correlation, more probability to have higher impact.

And in the modelling section, we studied and saw that RF and GBR where the best models when training or data set.

When comparing estimations of R2 score for the training set under LR, Lasso, SVM, RF, ABR and GBM algorithms, together with the performance metrics: MSE, RMSE and MAPE. Our discussion first starts with LR. Then we compared it with Lasso which obtained R2 score 0.4960990816544898. We saw that mean of CV score is 0.46607820211922485 which got my attention as it improved and got better than the scoring of LR which was 0.49827666044359065.

Then, we checked and compared the obtained results with algorithm Random Forest Regressor which got us R2 Score of 95.69% and a CV mean of 94.767%

Then we checked different algorithms like DT, KNN, SVR, Lasso, RFR,LR, AB.

Before deciding the best algorithm, we have gone through the several different steps such as: split data, train data through differents algorithms and their corresponding different parameters.

The list of each algorithm used is as following:

* + LinearRegression
  + DecisionTreeRegressor
  + KNeighborsRegressor
  + SVR
  + Lasso
  + RFR
  + AdaBoostRegressor
  + GradientBoostingRegressor

From which we selected GradientBoostingRegressor which gave me the best results in comparison with the other algorithms after applying GridSearchCV.

Following this is the estimation of Gradient Boosting Regressor. Employing alpha=0.098, learning\_rate=0.1, max\_depth=2, min\_samples\_leaf=2, n\_estimators=100, the results are then evaluated by the MAE, MSE and RMSE criteria. These values are estimated to be:

R2 score is: 0.9537134890052974

R2 score for train data: 0.965163437317112

Mean absolute error is: 400.82168588558903

Mean squared error is: 364747.4337817024

Root mean squared error is: 364747.4337817024

demonstrating that GBR is implying a reasonably good fit. In an iterative algorithm, the ‘optimal’ weight for each feature is indeed obtained by fine tuning an arbitrary value on a function, and going down its slope step by step until it reaches the lowest point of the function. Hence, our focus should concentrate on the explanatory power R2 and the error minimisation of the model.

Using random forest, our base model provides a R2 that is as high as 0.951524

for the test set while the MSE, RMSE and MAPE are estimated to be:

Mean absolute error is: 391.03356643356636

Mean squared error is: 382000.4956643356

Root mean squared error is: 382000.4956643356

In random forest, the estimation of SGD is not feasible because some of the hyperparameters are not continuous, such as the number of estimators. Rather, we can employ the grid search to minimise the model errors by tuning the hyperparameters.

For hyperparameter tuning, we perform many iterations of the entire fivefold cross-validation process, each of which sets different model settings, such as the number of estimators, min samples split, min sample leaf, max features, and max depth. Then, we compare all these models, pick up the best one, train it on the training set, and then evaluate on the test set as follows:



Hence, we take the result based on grid search as the best result associated with random forest.

If we do grid search with GBR instance that we created earlier, we have that:



We then obtain a R2R2 as high as 0.942737609028171 for the test set .The results based on grid search are the best results associated with gradient boosting regression. But we still will choose the previous

Lastly, we choose and check the results with the best parameters given by the GridSearch and we discuss the results based on the gradient boosting regression. In the base model, GBM provides a R2 of 0.9537134890052974 for the test set while the MSE, RMSE and MAPE are estimated to be:

Mean absolute error is: 400.82168588558903

Mean squared error is: 364747.4337817024

Root mean squared error is: 364747.4337817024

**Which is our FINAL SELECTION as ML algorithm for this case.**

It is very obvious that GBM fits our data reasonably well. In order to obtain even better results, we then employ the grid search to minimise the model errors by tuning the hyperparameters, including subsample, learning rates, the number of estimators, min samples split, min sample leaf, max features, and max depth:

For most of the time, GBM fits the data very well for example in this case. However, the model does not perform well with high bias for some extreme values that are lying far away from the cluster. Again, the model has achieved very good model fitting since most of the predicted values are lying close to the predicted values and its performance is slightly better than those of RF in relation to R2\_Score.

Finally, we compare the results of these estimation techniques. In terms of explanatory power, this study shows that RF and GBM have equally good performances. In terms of error minimisation, GBM is slightly better than RF with respect to the three performance error metrics and R2 score while it outperforms SVM. Hence, we demonstrate that GBM are very powerful tools for making accurate prediction of property prices, since the performances of this algorithm is higher to the other. In comparison, the performance of SVM is found to be below that of RF and GBM in all aspects. However, this does not imply that SVM is necessarily inferior.

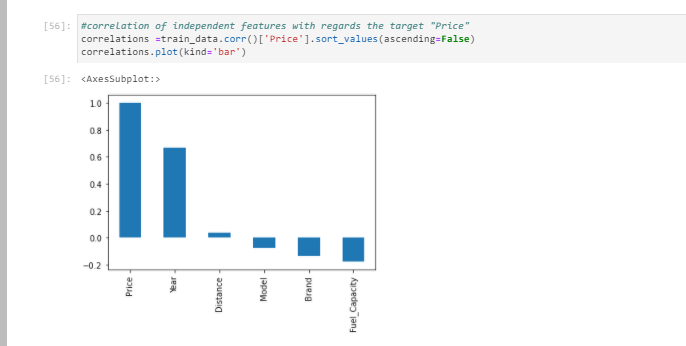
**CONCLUSION**

* Key Findings and Conclusions of the Study

If we concentrate on the main 3 questions, then the results answer the research questions as follows:

* Question 1 –Which variables are important to predict the price of variable?

If we first concentrate on the target and check the best correlations against it, then we will see that the following features are considered the best correlated features within our dataset:



Answer: We see several features such as Year, Fuel\_capacity shown in the above screenshot are the features that have higher impact comparing all the available features.

* Question 2 – Which machine learning algorithm performs better and has the most accurate result in used car price prediction? And why?

In general, the selection of a ML trained and tested model for a case like used car price prediction-estimation is very much influenced by the quality and availability of data. When we have wide range of data, the model can be trained easily and will be effective as sufficiently long time series of used car prices and relevant fundamental variables allow error-correction or vector error-correction models to be estimated.

So basically, here the study shows a comparison the regression algorithms work when predicting used car prices and eventhough RF gave the improved the results score compared to the results previously obtained, but GBR got the best R2 score and results overall. The final results of this study showed that GBR makes better prediction compared to other used algorithms.

Summarizing, the results answer the research questions as follows:

- Question 1 – Which machine learning algorithm performs better and has the most accurate result in used car price prediction? And why? GBR made the best performance overall when both R2 and RMSE scores are taking into consideration.

- Question 2 –Which variables are important to predict the price of variable? We see several features such as Year, Fuel\_capacity shown in the above screenshot are the features that have higher impact comparing all the available features.

- Question 3 - How do these variables describe the price of the used car?

These variables have higher impact on the target then the others when comparing the percentage of correlation. So, the higher correlation, more probability to have higher impact on the Price of the used car.

* Learning Outcomes of the Study in respect of Data Science

As we may know improvement in computer technology has made it possible data that cannot previously be captured, processed and analysed such as property research. This study is an attempt to use machine learning algorithms in estimating housing prices, and then compare their results.

In this study, our models are trained with housing property data using random forest (RF) and gradient boosting machine (GBM) among others. We have demonstrated that advanced machine learning algorithms can achieve very accurate prediction of property prices, as evaluated by the performance metrics. Given our dataset used for this study, our main conclusion is that RF and GBM are able to generate comparably accurate price estimations with lower prediction errors, compared with the SVM results.

The existing literature has demonstrated mixed results about which algorithm is superior in making predictions. Previous studies suggest that RF has a better performance than GBR but other studies demonstrate that both models have comparable predictive power and almost equally applicable in making predictions (Ahmad et al., [2017](https://www.tandfonline.com/doi/full/10.1080/09599916.2020.1832558)). In another study about the forecast of daily lake water levels, RF is found to exhibit the best results when compared to linear regression, SVM and ANN with respect to R2R2 and RMSE (Li et al., [2016](https://www.tandfonline.com/doi/full/10.1080/09599916.2020.1832558)). Utilising RF, stochastic gradient boosting and SVM to predict genomic breeding values, Ogutu et al. ([2011](https://www.tandfonline.com/doi/full/10.1080/09599916.2020.1832558)) conclude that boosting has had the best performance, followed by SVM and RF. In this study, SVM performs better than RF (see also Caruana & Niculescu-Mizil, [2006](https://www.tandfonline.com/doi/full/10.1080/09599916.2020.1832558)).

I think we should suggest that the choice of algorithm depends on several factors including, the size of data set, computing power, time constraint and researcher’s knowledge about machine learning. If the accuracy of prediction is the priority, we recommend utilising RF or GBM, although their computations often take much longer times than SVM. Needless to say, it is a good practice to use more than one algorithm to compare the predictions.

* Limitations of this work and Scope for Future Work

To conclude, the application of machine learning in car used research is still at an early stage. We hope this study can provide some methodological improvements and contributions to used car calculation and presenting an alternative to the valuation of used car prices. Future direction of research may consider incorporating additional used car transaction data including more features.

Even though, we are very happy with our final model’s performance, but here we can improve the results by working more intensively on the model and performance results. The inclusion of GBR for continuous variables improved our model’s performance. That being said, there are still additional features and methods we would like to try that we think could further improve the model’s performance, as we already mentioned.

Future work on this study could be divided into main areas to improve the result even further. Which can be done by:

- The used pre-processing methods do help in the prediction accuracy. So, experimenting with different combinations of pre-processing methods to achieve better prediction accuracy.

- Make use of the available features and if they could be combined as binning features that can improve the data and hence the performance of the model and their results.

- Training the datasets with different regression methods such as Elastic net regression that combines both L1 and L2 norms. In order to expand the comparison and check the performance.

- Some of the factors that have been studied and analised in this study has a weak correlation with the used car price. Hence, by adding more factors to the dataset that affect the used car location, engine characteristics and the average rate of similar used cars in order to increase the number of factors that have an impact on used car prices. This could also lead to a better finding for question 1 and 2, hence also 3.

. - The results of this study have shown that GBR is prone to overfitting. However, GBR still a strong algorithm that has a lot of options that could, with the right methods, provide a better prediction accuracy.

For future work, we recommend that working on large dataset would yield a better and real picture about the model training and testing results. We have undertaken only few Machine Learning algorithms that are regressors, but we need to train many other regressors and understand their predicting behaviour for continuous values. By improving the error values this work can be useful for development of applications for various respective used car sale pricing.