

Car Price Prediction

Submitted by

Sharanya Susheel

**ACKNOWLEDGMENT**

Firstly, would like to thank Flip Robo for giving me this opportunity and this learning will stay with me for a life time , special thanks to my mentors and career coach at Data Trained who have been my constant source of support and last but not the least would like to thank my family without whose support I would ‘nt be able to finish this project.

Below are the references which helped me in competing my project:

<https://colab.research.google>.

<https://www.temjournal.com/content/81/TEMJournalFebruary2019_113_118.pdf>

https://www.researchgate.net/publication/343878698\_Used\_Cars\_Price\_Prediction\_using\_Supervised\_Learning\_Techniques

**INTRODUCTION**

* Business Problem Framing

Its always fascinating to see how we could predict things better when we have large sums of historic data. The world has advanced a lot in terms of technology, medicine, education and many other verticals with the right use of historic data.

One such vertical where it can make a huge impact is the car industry, as the main reasons are, highly fluctuating price which depends on expected booking, actual booking, events and a lot more factors. Now that we all know some basics like the weekends are expensive and holiday season its expensive, we actually don’t know the best price we can book at any given point of the time in the year. Imagine if we could get alerts and recommendations when the price is actually low or offered the best. It does save a lot of money for the travelers and moreover it can also help regulate the demand and availability for business consistency with the airline companies. And it is not mere historic data evaluation done by us, which becomes cumbersome after a point in time with large sums of data, it is how we write algorithms for machines to process and learn large sums of data to help us achieve business objective, which simply is called machine learning.

Using regression method in machine learning, it has helped us overcome the said concern. We need to identify the data and time variables, missing values, heat map to help us create data frame, which further will help us build our model. We have also used RandomForestRegressor, LinearRegression, KNeighborsRegressor and DecisionTreeRegresssor, cross validation using randomized search CV for achieving the results.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

This problem is about car price prediction. The sample dataset is obtained by web scraping from various websites. With the covid 19 impact in the market, we have seen lot of changes in the car market. Now some cars are in demand hence making them costly and some are not in demand hence cheaper. One of our clients works with small traders, who sell used cars. With the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models.

This data set contains 301 rows and 9 columns.

The project has been divided into two parts,

Part 1

Exploratory Data Analysis and Data Cleaning :

Part 2:

Training a machine learning model

* Data Sources and their formats
* As told earlier the data is obtained by web scraping .The data is then copied into excel,converted into csv and then used to build ourmodel.
* Data Preprocessing Done
* The steps followed for data cleaning are described as follows:
* 1)Firstly the information about the dataset was checked and it gave the data shape i.e the total number of rows and columns.
* 2)The datatypes of all the columns were checked to find out if they are object, integer or float.
* 3)Duplicate rows were dropped in the dataset
* 4)Null values were checked using df.isnull().sum() function.
* 5) After that we check the summary statistics of our dataset. This part tells about the statistics of our dataset i.e. mean, median, max value ,min values and also it tell whether outliers are present in our dataset or not.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)
* It is a regression type of problem.Linear Regression,Random Forest Regressor,Descion tree regressor etc were used.
* Testing of Identified Approaches (Algorithms)
* Lg=LinearRegression()
* Rfr=RandomForest Regressor()
* Dtr=DecisionTreeRegressor()
* Hyper tuned ridge
* Hyper tuned lasso
* Run and Evaluate selected models

Linear regression- LinearRegression fits a linear model with coefficients w = (w1, …, wp) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.

lg = LinearRegression()

results(lg,X\_train,X\_test,y\_train,y\_test)

RandomForest Regressor- A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max\_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

rfr = RandomForestRegressor()

results(rfr,X\_train,X\_test,y\_train,y\_test)

RandomForestRegressor()

dtr =DecisionTreeRegressor()

results(dtr,X\_train,X\_test,y\_train,y\_test)

Hyper tuned ridge- This model solves a regression model where the loss function is the linear least squares function and regularization is given by the l2-norm. Also known as Ridge Regression or Tikhonov regularization. This estimator has built-in support for multi-variate regression (i.e., when y is a 2d-array of shape (n\_samples, n\_targets)).

params = {"alpha": [.01, .1, .5, .7, 1, 1.5, 2, 2.5, 3, 5, 8,9,10,11,12,13,15,16,17,18,19,20,21,22,23,24,25,26,27,30]}

ridreg = Ridge()

clf = GridSearchCV(estimator=ridreg, param\_grid=params, cv=5, return\_train\_score=True)

clf.fit(X\_train, y\_train)

## Hyper tuned Lasso- Lasso

[Lasso](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html#sklearn.linear_model.Lasso) does a similar thing but uses the sum of absolute values (the l1 norm) of the weights for the penalty.

*Note: there is also a n\_samples in the sklearn formula, which is the number of observations, which should not change for the same X and y. I have found no explanation as to why that is there, perhaps to compare different models, please let me know if you have a better idea.*

Text

Description automatically generated with medium confidence

Lasso formula

Lasso will start decreasing the coefficients of variables that are not so important, and it is also possible to decrease coefficients down to 0. In layman’s terms:

*X1, your minimal contribution to the overall score is noted. However, in light of the latest penalty scores, we will have to let you go from the regression. It’s just not worth keeping you around.*

params = {"alpha": [.00001, .0001, .001, .005, .01, .1, 1, 5]}

lasreg = Lasso()

clf = GridSearchCV(estimator=lasreg, param\_grid=params, cv=5, return\_train\_score=True)

clf.fit(X\_train, y\_train)

Visualizations

Shape, square

Description automatically generated

This heat map shows the absence of null values in our dataset.

Calendar

Description automatically generated with low confidence

This heat map shows that 'Selling\_Price' and 'Kms\_Driven' are highly correlated with each other.

A picture containing diagram

Description automatically generated

This is a collective representation of outliers in the different columns.

Individual boxplot for the representation of outliers.

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

A picture containing chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Graphical representation of skewness

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

A picture containing chart

Description automatically generated

Logo, icon

Description automatically generated

from the above graphs we can observe the following: 1)The number of dealer cars is higher than the individual cars 2)The number of manual cars is higher than that ofautomatic cars 3)The number of the third party owners is comparitvely low 4)The number of cars with CNG type of fues is very low compared to that with petrol and diesel.

Chart, box and whisker chart

Description automatically generated

From the above graphs we can observe that the catergorical data has some sort of pattern regarding car's present price and selling price.

Chart, histogram

Description automatically generated

Car usage in years andits selling price has a negative correlation

so we can say that as many years car gets used the value decreases.

Present price and selling price are positively correlated

From the graph we can also see that the selling price is not decreasing gradually except in one case.

Linear Regression

Chart, histogram

Description automatically generated

RandomForest Regressor

A picture containing text, wall

Description automatically generated

Decision Tree Regressor

A picture containing text, wall

Description automatically generated

Hyper tuned ridge

Chart, histogram

Description automatically generated

Hyper tuned lasso

Chart, histogram

Description automatically generated

* Interpretation of the Results

| **model** | **r^2** | **cv\_score** | **mae** | **mse** |
| --- | --- | --- | --- | --- |
| **0** | linear | 0.873934 | 3.747733 | 1.023575 | 2.878199 |
| **1** | random Forest | 0.901817 | 2.874695 | 0.626741 | 2.241609 |
| **2** | Dicision Tree | 0.864997 | 3.738119 | 0.924933 | 3.082233 |
| **3** | Ridge | 0.880716 | 3.610160 | 0.975591 | 2.723360 |
| **4** | Lasso | 0.882420 | 3.957253 | 0.982370 | 2.684460 |

**CONCLUSION**

* Key Findings and Conclusions of the Study
* Random forest has a high r2 score which is a better model compared to the others.