

**Car Price Project**

Submitted by:

AKASH KRISHNA MAURYA

**ACKNOWLEDGMENT**

* I would like to thank my SME Miss Khushboo Garg for giving me this golden opportunity to work on this project, used car price prediction.
* Some of the articles and research papers that I find useful for completion of this project are:

**INTRODUCTION**

**Business Problem Framing:**

* 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.
* They are looking for new machine learning models from new data. We have to make car price valuation model.

**Conceptual Background of the Domain Problem:**

* In this project, we scraped data of used cars from different website like olx, cars24, droom etc. After that we will build a machine learning model to predict the price of used cars.
* This model will then be used by our client to understand how exactly the used car prices vary after covid-19. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns.

**Review of Literature:**

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

**Motivation for the Problem Undertaken:**

As we know that used car industries market is really huge, data science can contribute majorly in this industry as there are numerous industries working in this domain. By building a model for the car price prediction, it can help the industries to plan their strategies accordingly, and it can help them to mould themselves as the future trend goes by.

**Analytical Problem Framing**

**Mathematical/ Analytical Modeling of the Problem:**

* **Standard-Scaler:** After removing skewness, we need to scale our data. For this we use standard scaler method. This method normalizes our data and essential for machine learning algorithms that calculate distance between data. For instance, most of the classifiers calculate the distance between two points by the distance. If one of the features has large value, then distance consider that particular feature. This method is necessary, where large and small values present in our data. This method transforms our data with mean = 0 and standard deviation = 1.
* **Label Encoder:** Label encoder is one method of converting data to prepare it for an algorithm and get a better prediction. With label encoding, we convert each categorical value into a new categorical column and assign a value starting from 0.Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.
* **Z-score:** Outliers are extreme values that fall a long way outside of the other observations. In this project we used z-score method to remove outliers. In this procedure we calculate the z-score for each observation. Any z-score greater than 3 or less than -3 is considered to be an outlier. This rule of thumb is based on the empirical rule. From this rule we see that almost all of the data (99.7%) should be within three standard deviations from the mean. To calculate the z-score, we subtract the mean from the data point, and then divide by our standard deviation.

**Data Sources and their formats:**

We scraped the data from olx, cars24, droom.com etc. The data contains features like Brand name, model name, location of the car, manufacture year of the car, number of owners that the car had previously, fuel used for the car, gear (manual or automatic), driven (how much the car was driven before) and the reselling price.

**Data -Preprocessing Done:**

* Firstly, we checked for the null values, there were 300 null values present in the model feature.But when I scraped the data there were dashes in the data.
* Then we checked for the dashes present in the dataset, pretty much every column had null values.
* We imputed all the null values with their mode, and removed the null values.
* Then we had to remove the extra information present in the columns.
* In the price column we removed the rupee symbol and the commas.
* We removed the KM sign from the driven\_kms column.
* Now we had to convert these columns into float datatypes, as before we checked it was object datatype.
* We converted the driven\_lkm and car price column into float datatype.
* In the owner’s column when we checked the unique values present, there were unnecessary noises present in the column, like 1st and first, they both indicate the same thing so we converted the string first to 1st and likewise other values too.
* There were noises present in the manufacture year too, for example 2015 and 2015.0, they both indicate the same thing, hence we converted the 2015.0 to 2015 and we converted the other values in the same manner.
* After removal of the null values and noises, we converted the object datatype to numerical datatype as it is important for machine learning.
* We used label encoder to encode all the columns.
* Then we checked for the outliers with the help of zscore method, removed the outliers, the data loss was just 6%.
* Then with the help of standard scaler we scaled the dataset in one scale, so that there is a better accuracy score and easy machine learning process.
* Finally, we separated the x and y variables, x being the feature columns and y being the target variable.

**Data Inputs- Logic- Output Relationships:**

* We used regression machine learning models because our target variable is car price. So, we need to find price of used cars.
* There are many regression models but here we used some of them models.
* First, we split our training dataset into two segments: training and testing. We take 67% data for training and 33% data for testing. For splitting data, we use train test split method. Below is the code for splitting the data:



1) 85% of the observation as training set-->x\_train

2) The associated target for each observation in x\_train -->y\_train

3) 15% of the observation as test set-->x\_test

4) The target associated with the test set-->y\_test.

* After splitting data, we passed training data to machine learning models. The fitted model will first be used to generate prediction on the test set (x\_test). Next, the predicted class labels are compared to the actual observed class label (y\_test) to see the difference between them.

**Hardware and Software Requirements and Tools Used:**

Hardware Requirements: -

1. Processor: 7th gen core i5 or above.
2. RAM: 4 GB DDR3 or above.
3. HDD/SDD: 128 GB or above.
4. GPU: Intel iris plus graphics 640 1536 MB or above.

Software Requirements: -

Anaconda software must be installed with all the necessary libraries like pandas, numpy, matplotlib, seaborn, scipy, sklearn.

**Model/s Development and Evaluation**

**Identification of possible problem-solving approaches (methods):**

* Data reading and understanding
* Data cleaning
* Data analysis
* Handling outliers
* Handling skewness
* Encoding data
* Scaling
* Train test split
* Machine learning algorithms

**Testing of Identified Approaches (Algorithms):**

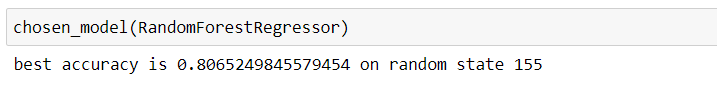
* Random Forest Regressor.
* Decision Tree Regressor.
* Ada Boost Regressor.
* Gradient Boosting Regressor.
* Support vector Regressor.
* Lasso
* Ridge
* Linear Regression

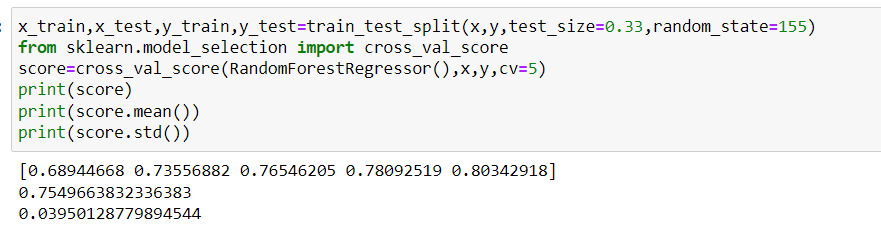
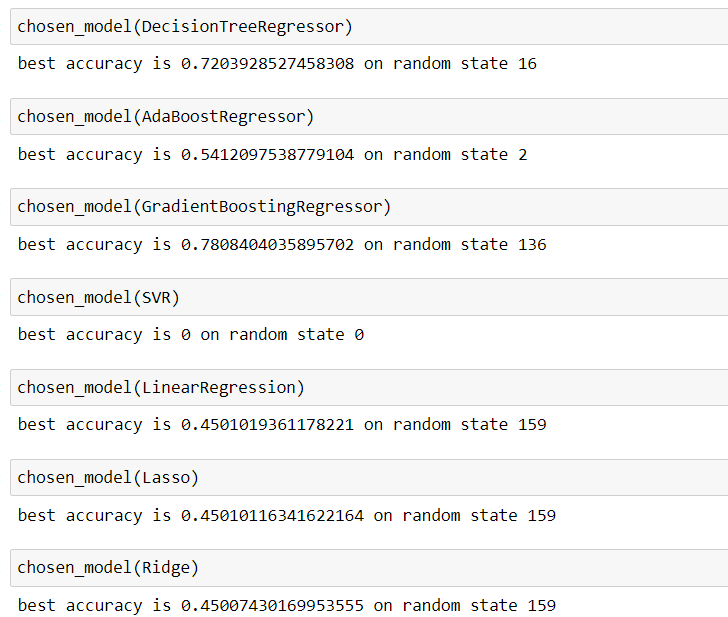
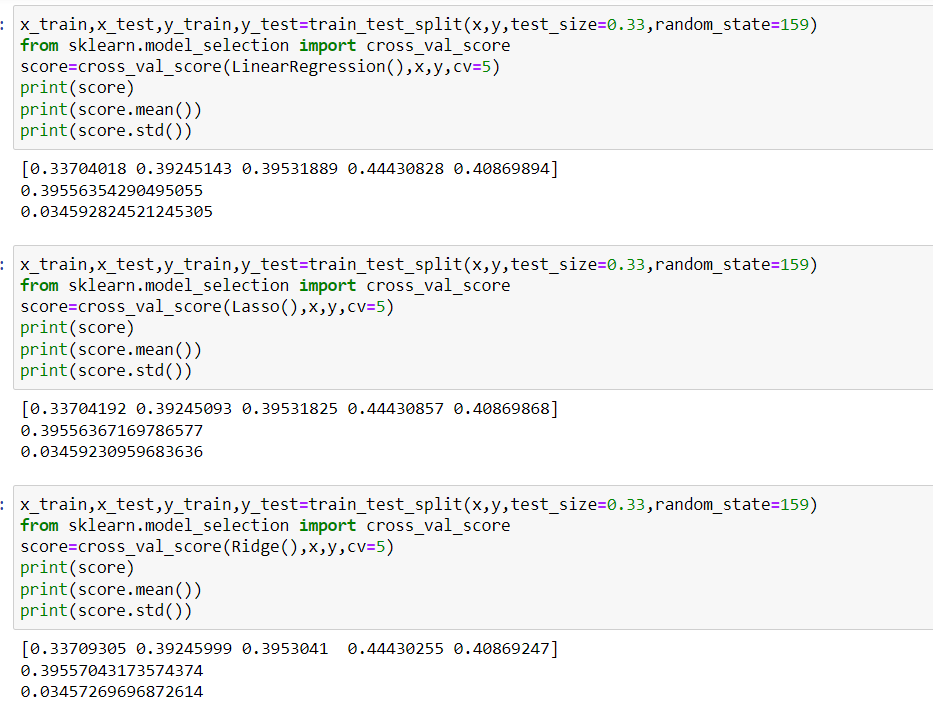
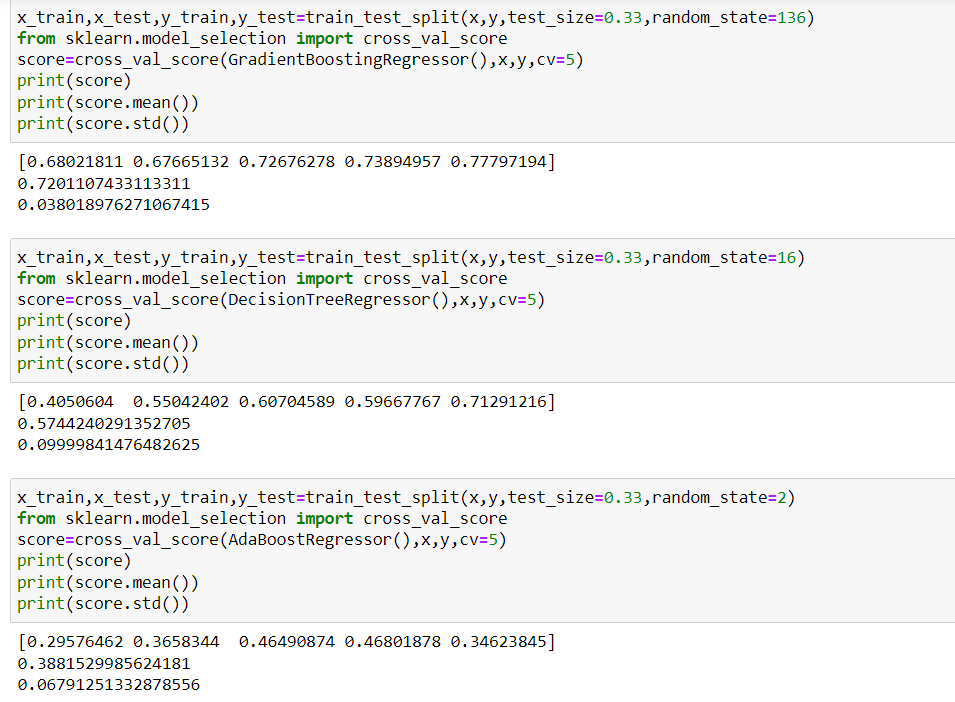
**Run and Evaluate selected models:**

* **Random forest 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 themax\_samples parameter if bootstrap=true (default), otherwise the whole dataset is used to build each tree.

Above snapshot is of the code, where I have made a function chosen model, which trains the model and gives the R2\_score in the best random state.



* For random forest regressor it’s giving the best accuracy score 80.65% at 155 random state.
* Now we will check for the overfitting or underfitting of the model.
* From the above code we can deduce there is a bit of overfitting in the model. Here we took cv=5, which indicates the training dataset is divided into 5 parts, and mean of those 5 parts gives us the cross-value score.
* As this is our best model so we went forward with this model, and did the hyperparameter tuning.
* We chose the parameters and trained our model, and after hyperparameter tuning 77.32% is our r2\_score.
* Similarly checked the score for other algorithms too.
* Below are the r2 scores of each model in the best random state.
* We checked for other scores also like cross-value score, mean absolute error and root mean squared error of all the models. After careful consideration of each metrics we chose random forest as our best model.**Key Metrics for success in solving problem under consideration:**
* R2\_score: R2 score is the percentage of variation explained by the relationship between two variables. Range of the r2 score is varies from 0 to 1. Mathematical formula of the r2 score is as below:

R 2= 1- SSres / SStot

Where, SSres is the sum of squares of the residual errors. SStot is the total sum of the errors.

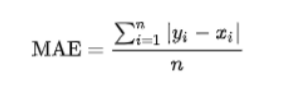
Mean absolute error (MAE): The Mean Absolute Error, also known as MAE, is one of the many metrics for summarizing and assessing the quality of a machine learning model.

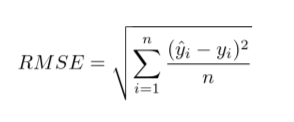
* Mean absolute error subtract the predicted value from actual value as below:

Prediction Error → Actual Value - Predicted Value

* This prediction error is taking for each record after which we convert all error to positive. This is achieved by taking Absolute value for each error as below:

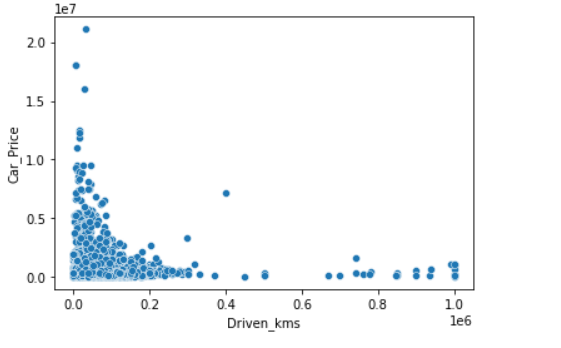
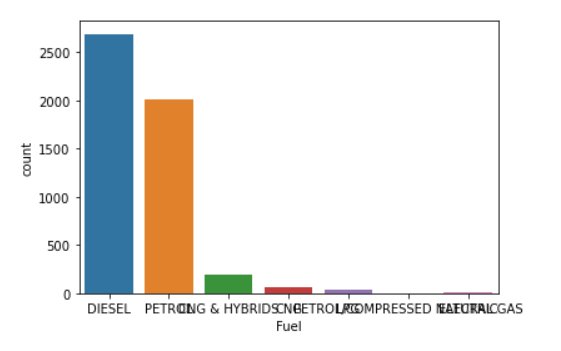
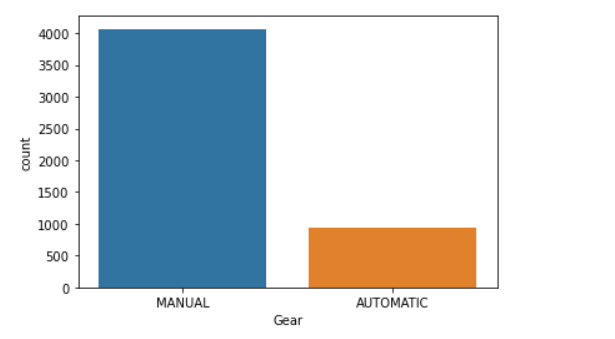
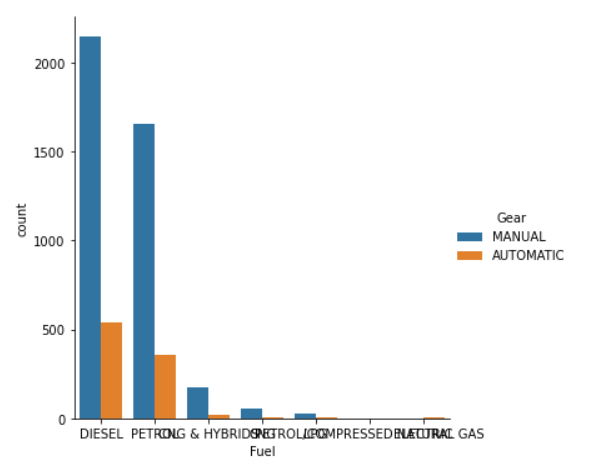
Absolute Error → |Prediction Error|

* Finally, we calculate the mean for all recorded absolute errors (Average sum of all absolute errors). Below is the formula of MAE:
* Root mean squared error (RMSE): Root Mean Square Error is the measure of how well a regression line fits the data points. RMSE can also be construed as Standard Deviation in the residuals.
* It is a standard way to measure the error of a model in predicting quantitative data. Formally it is defined as follows:



* ŷ1, ŷ2,… ŷn are predicted values
* y1,y2,…yn are observed values
* n is the number of observations

**Visualizations:**

* Firstly we saw the scatter plot between the car price and car driven, we deduced from the plot, that more the car is driven in the past, less will be the price of car, as it clearly indicates that the car is more used, less will be the price.
* This countplot suggest that maximum vehicles which are getting resell have diesel as the fuel. This also tells us about there might be a common problem with diesel vehicles.
* This countplot suggest that maximum vehicles have manual as the gear. It also suggests that in India maximum population have manual cars.
* This catplot shows that diesel manual cars are more on resell.

**Interpretation of the Results:**

* After seeing the visualizations, we saw that there was a direct relationship between the price of the car and how much the car was driven, more the car was driven, less will be the price of the car.
* We also saw that diesel manual cars were more on the list of resell.
* And the count of Maruti car were more than the other brands.
* We had to remove all the unwanted things from the data and convert it into readable format for machine learning purpose.
* We imputed all the null values and made better for ML purpose.
* After seeing all the metrics, we chose Random Forest regressor as our best model.

**CONCLUSION**

**Key Findings and Conclusions of the Study:**

* The purpose of this article was twofold: to understand the pattern of used cars market and make predictive model, which is able to effectively predict the price of used cars.
* We use many algorithms to find best model and best result were observed of the random forest regressor with 78% r2 score accuracy with good mean absolute error.
* There are many variables important to predict the price of houses. Like driven kilometres, car model, fuel of car etc.
* By using machine learning model our client can decide whether to increase or decrease the price of used cars.

**Limitations of this work and Scope for Future Work:**

* In future we may add large historical data of car price which can help to improve accuracy of the machine learning model.
* For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset.