

**Field of Study**: Artificial intelligence

**Project Name**: Car Price Prediction

Group Member

* Niloy Saha - 183-15-12027
* Md. Mushiur Rahman - 183-15-11846
* S.M. Towhidul Islam - 183-15-11970
* Apurba Roy - 183-15-11850

Submitted To

Md. Tarek Habib

Assistant Professor

Daffodil International University

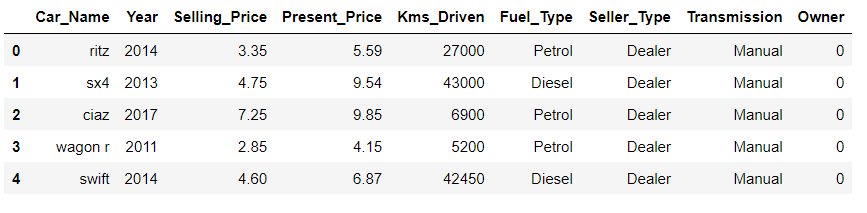
**Abstract**

Determining whether the listed price of a used car is a challenging task, due to the many factors that drive a used vehicle’s price on the market. The focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features, in order to make informed purchases. We implement and evaluate various learning methods on a dataset consisting of the sale prices of different makes and models across India. Our results show that Random Forest model yield the best results. Conventional linear regression also yielded satisfactory results, with the advantage of a significantly lower training time in comparison to the aforementioned methods.

Problem statement

We have to predict at what price a particular used car can be sold in other words using the dataset or the dataset feature’s we need to train our machine learning algorithm to find a pattern so that once we give new information it can deliver us an estimated value for selling the car.

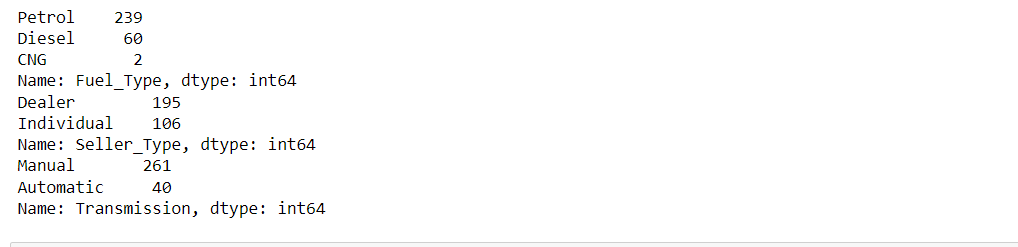
Dataset

For this project, we are using the dataset on used car sales from India, available on Kaggle [1]. The features available in this dataset are Mileage, Brand, Year, Sold Price, KMS Driven, Fuel Type, Seller Type, Transmission Type, Owners. 

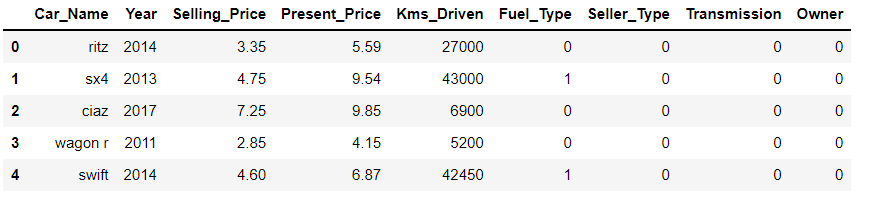
Pre-processing

In order to get a better understanding of the data, we noticed that there are 301 values of different cars so what we’re basically trying to do is train our machine learning model with all these selling prices and other features of the car and once it has learned from this data then we can give new information to the model so that the machine learning model can predict how much price the car can be sold

One another thing to notice if the car is very old, the price will be somewhat less and if the kilometers driven is very high then a person won’t buy it for a high price so these are several features on which the price depends

We have also checked the number of values for this categorical data like follows: -

We have encoded our data which’s important because machine needs 0 or 1 not some plain text data so what we have done the processing for Fuel | Seller | Transmission for example if the fuel type is petrol (0) and Diesel (1) and CNG (2) respectively in Boolean category.



Analyzing Linearity in Dataset

To analyze the degree to which our features are linearly related to price, we plotted the Price against Year for a particular Make and Model. There seemed to be a fair degree of linearity which’s shown below: -

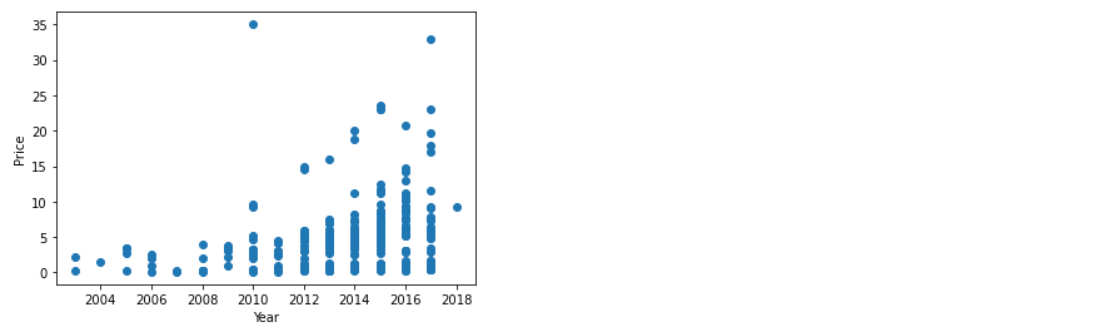
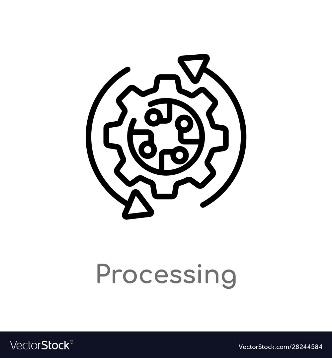
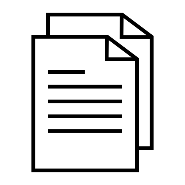


Fig: price vs year scatter plot

Workflow





**Car data** **Train Test split**

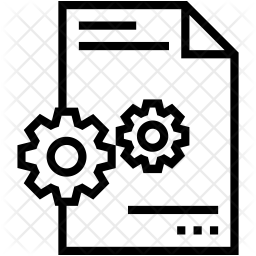
**machine learning model**







**Trained model** Car Price

**New data prediction** 

Methodology

We utilized several classic and state-of-the-art methods, including ensemble learning techniques, with a 90% - 10% split for the training and test data. For most of the model implementations, the open-source Scikit-Learn package was used.

**1. Linear Regression**

Linear Regression was chosen as the first model due to its simplicity and comparatively small training time. No regularization was used since the results clearly showed low variance

**2. Lesso Regression**

Lasso is a modification of linear regression, where the model is penalized for the sum of absolute values of the weights. The value of R-square for our model has been increased using this model.

**3. Random Forest**

Random Forest is an ensemble learning based regression model. It uses a model called decision tree, specifically as the name suggests, multiple decision trees to generate the ensemble model which collectively produces a prediction. The benefit of this model is that the trees are produced in parallel and are relatively uncorrelated, this model was hence chosen to account for the large number of features in the dataset and compare a bagging technique with the following gradient boosting methods

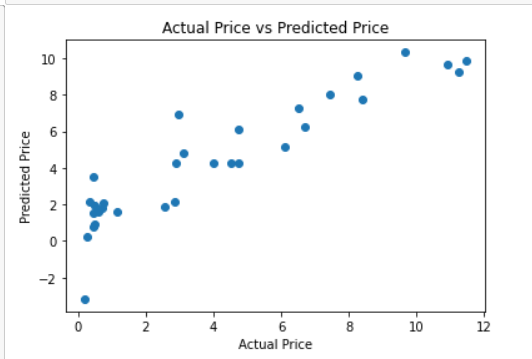
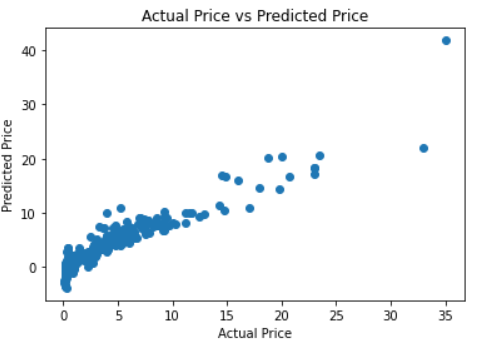
**Result**

The results of our tests were quantified in terms of the R score of our predictions. R score is a statistical measure of how close the data are to the fitted regression line.

|  |  |  |
| --- | --- | --- |
| Learning Algorithm | R square score on Test data | R square score on Train data |
| Linear Regression | 0.83 | 0.87 |
| Lasso Regression | 0.87 | 0.84 |
| Random Forest | 0.98 | 0.98 |

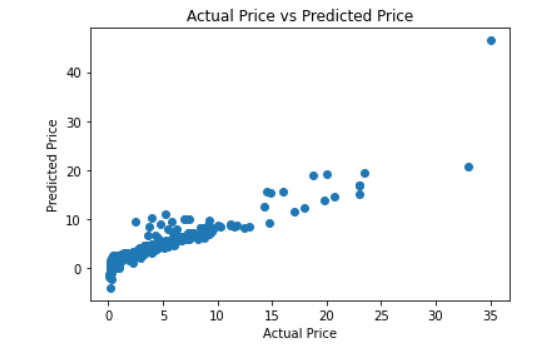
Clearly we can see Random forest performing well. This can be attributed to the apparent linearity of the dataset. We believe that it can also be attributed to the difficulty in tuning the hyper parameters for most gradient boost methods

Linear Regression Model



Train Data Test Data

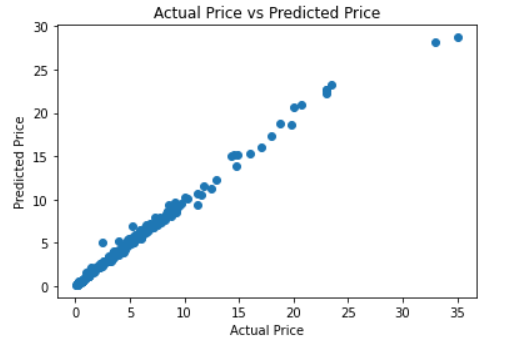
Lasso Regression Model



Train Data Test Data

Random Forest





Train Data Test Data

**Future Work**

For better performance, we are planning to test or train these data in models like LIGHT GBM, XGBOOST, Gradient Boost and Deep Neural Network to see the variance and also use adaptive learning rates to see the linearity in more efficient manner.

**Contributions**

All team members contributed equally towards this project.

Link to project repository: <https://github.com/newlyyy/car-price-prediction>