FLIP ROBO

Car Price Prediction Project

Submitted by:

Rutuja Patil

ACKNOWLEDGMENT

Because of new computing technologies, machine learning today is not like machine learning of the past. It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks; researchers interested in artificial intelligence wanted to see if computers could learn from data. The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results. It’s a science that’s not new – but one that has gained fresh momentum. While there is an end number of applications of machine learning in real life one of the most prominent application is the prediction problems. There are various topics on which the prediction can be applied. One such application is what this project is focused upon. Websites recommending items you might like based on previous purchases are using machine learning to analyze your buying history – and promote other items you'd be interested in. This ability to capture data, analyze it and use it to personalize a shopping experience (or implement a marketing campaign) is the future of retail.

INTRODUCTION

- Business Problem Framing

From a long time since being, a continuous paradigm of transactions of commodities has been into existence. Earlier these transactions were in the form of barter system which later was translated into a monetary system. And with consideration into these, all changes that were brought about the pattern of re-selling items was affected as well. There are two ways in which the re-selling of the item is carried out. One is offline and the other being online. In offline transactions, there is a mediator present in between who is very vulnerable to being corrupt and make overly profitable transactions. The second option is online wherein there is a certain platform which lets the user find the price he might get if he goes for selling.

a. Kilometers traveled – We know that the number of kilometers traveled by a vehicle has a huge role to play while putting the vehicle up for sale. The more the vehicle has traveled, the older it is.
b. Year of registration – It is the year when the vehicle was registered with the Road Transport Authority. The newer the vehicle is; the better value it will yield. By every passing year, the value will depreciate.
c. Fuel Type – There were two types of fuel types present in the dataset that we had. Petrol and Diesel. It was relatively less dominant.

It's due to the above factors that we need a system that can develop a self-learning machine learning-based system. This was the basis on which a set of objectives was supposed to be formulated. One thing that was pre-determined was that this is going to be a real-time project.

- Conceptual Background of the Domain Problem

1) To build a supervised machine learning model for forecasting value of a vehicle based on multiple attributes.
2) The system that is being built must be feature based i.e. feature wise prediction must be possible

- ## Review of Literature

In this project, we discuss various applications and methods which inspired us to build our project. We did a background survey regarding the basic ideas of our project and used those ideas for the collection of information like the technological stack, algorithms, and shortcomings of our project which led us to build a better project.

1)CARS24

Cars24 is a web platform where seller can sell their used car. It is an Indian Start-up with a simplified user interface which asks seller parameters like car model, kilometers traveled, year of registration and vehicle type (petrol, diesel). These allow the web model to run certain algorithms on given parameters and predict the price.

2)GET VEHICLE PRICE-

Get Vehicle Price is an android app which works on similar parameters as of Cars24. This app predicts vehicle prices on various parameter like Fiscal power, horsepower, kilometers traveled. This app uses a machine learning approach to predict the price of a car, bike, electric vehicle and hybrid vehicle. This app can predict the price of any vehicle because of the smartly optimized algorithm.

- ## Motivation for the Problem Undertaken

The automotive industry is composed of a few top global multinational players and several retailers. The multinational players are mainly manufacturers by trade whereas the retail market features players who deal in both new and used vehicles. The used car market has demonstrated a significant growth in value contributing to the larger share of the overall market. The used car market in India accounts for nearly 3.4 million vehicles per year.

FEATURES-

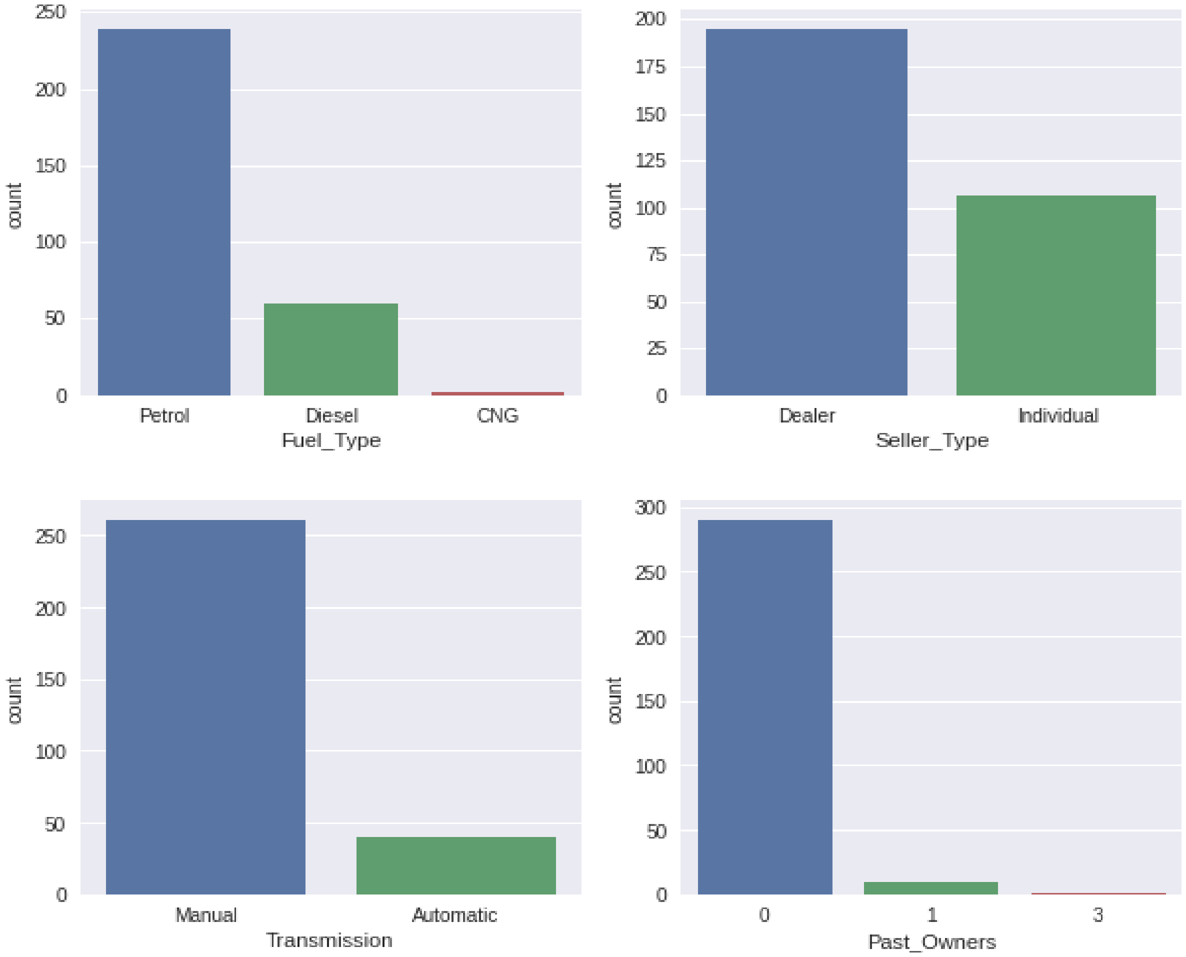
There will be majorly two features provided in the project note that this will be not

- Re-sale platform: A centralized platform for car re_sale that will predict prices.
- Feature selection: Feature-based search and prediction.

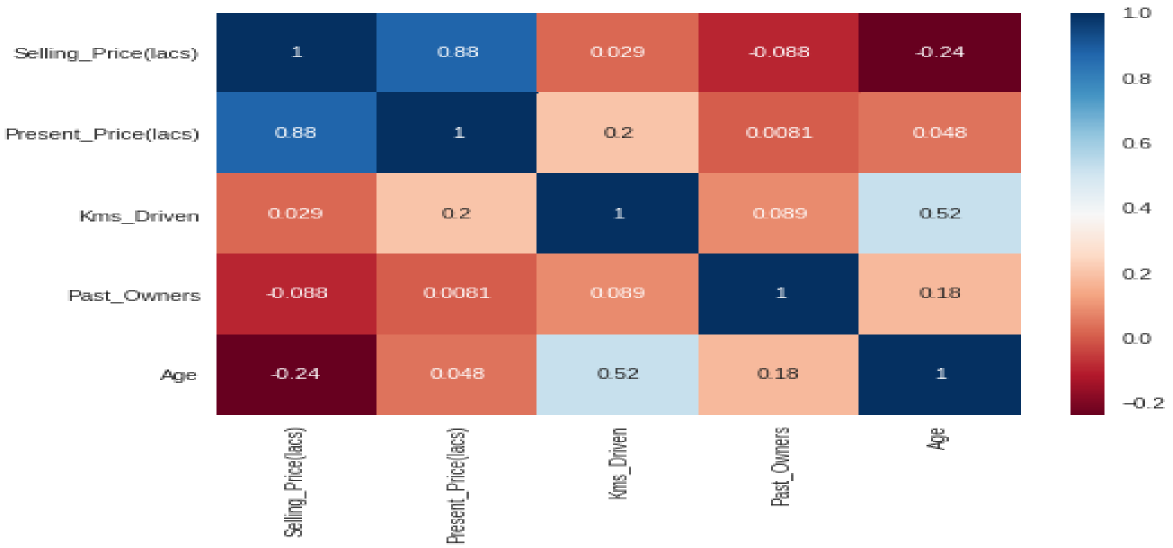
Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

1)Univariate Analysis-



2)Multi variate Analysis-



- ## Data Sources and their formats

  For this project, we are using the dataset on used car sales from a data analysis, Regression algorithms.

- ## Data Preprocessing Done

  In this section we plotted our data in two steps-

  1)creating dummies for categorical features

  2)Train-test split

- ## Hardware and Software Requirements and Tools Used

  Python was the major technology used for the implementation of machine learning concepts the reason being that there are numerous inbuilt methods in the form of packaged libraries present in python. Following are prominent libraries/tools we used in our project.

  1)NUMPY

  NumPy is a general-purpose array-processing package. it provides a high-performance multidimensional array object and tools for working with these arrays. It is the fundamental package for scientific computing with Python. Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data.

  Arbitrary data-types can be defined using Numpy which allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

  3)JUPYTER NOTEBOOK

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. It includes data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  1)Reading and Understanding the dataset.

  2)Data Preprocessing.

  3) Exploratory Data Analysis

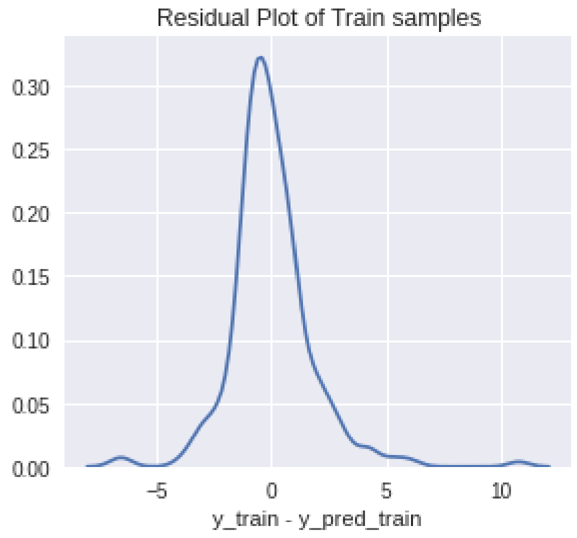
  4)Model creation and Evaluation

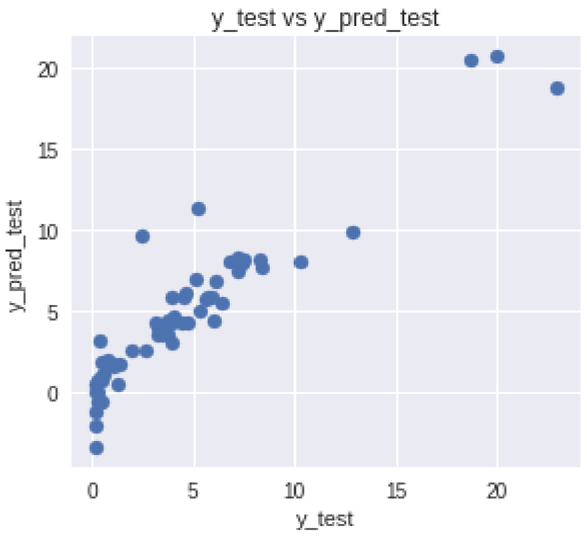
- Testing of Identified Approaches (Algorithms)

  1. Linear Regression
  2. Ridge Regression
  3. Lasso Regression
  4. Random Forest Regression
  5. Gradient Boosting regression

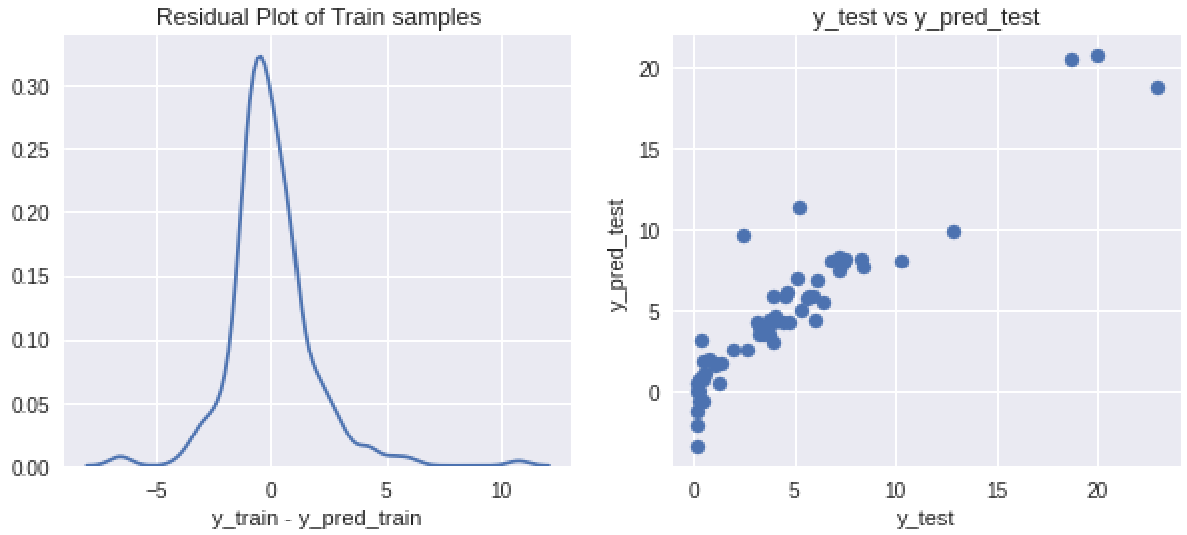
1)Linear Regression-

Linear Regression was chosen as the first model due to its simplicity and comparatively small training time. The features, without any feature mapping, were used directly as the feature vectors.No regularization was used since the results clearly showed low variance.

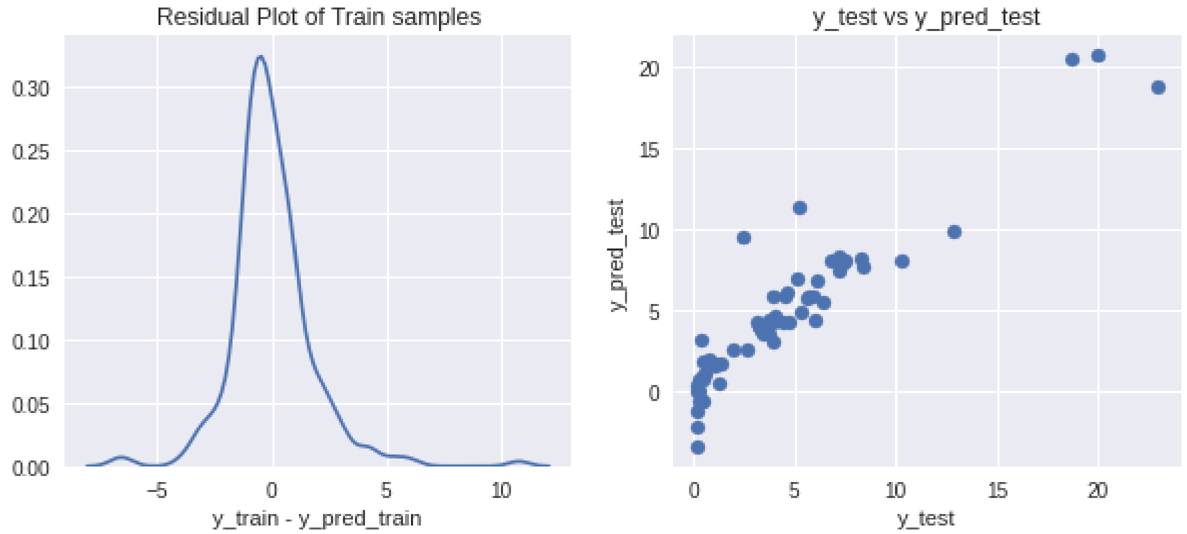




2)Ridge Regression-

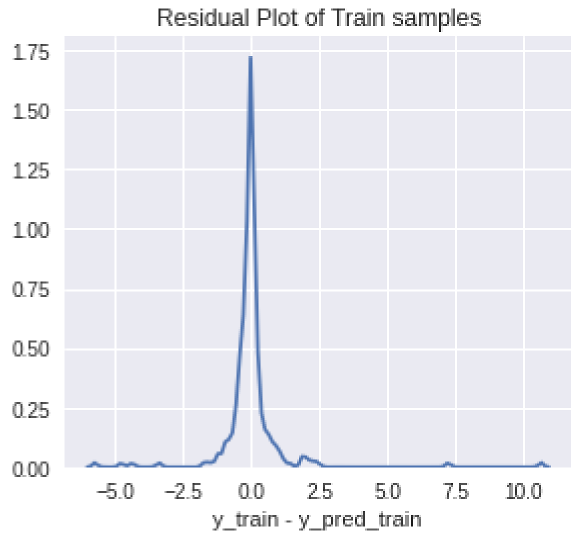


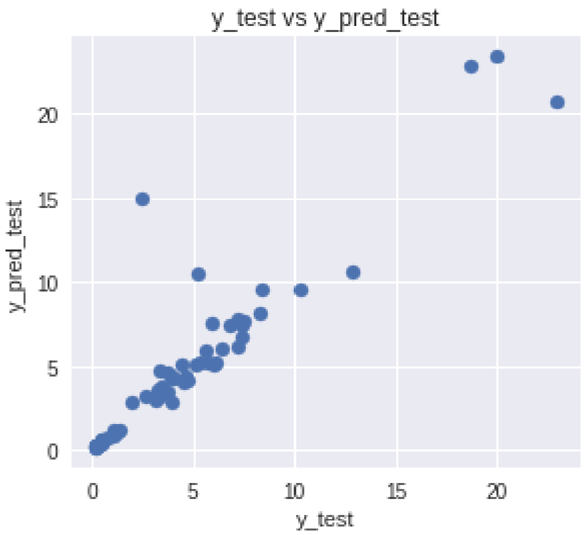
3)Lasso Regression-



4)Random Forest Regreesion-

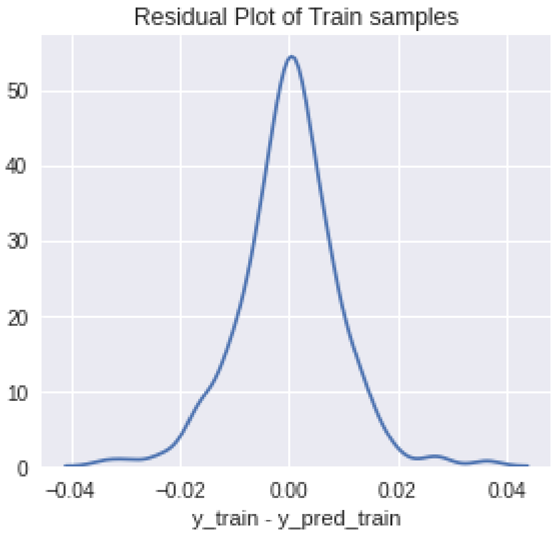
Random Forest is an ensemble learning based regression model. It uses a model called decision tree, specifically as the name suggests, multiple decision trees to generate the ensemble model which collectively produces a prediction. The benefit of this model is that the trees are produced in parallel and are relatively uncorrelated, thus producing good results as each tree is not prone to individual errors of other trees. This uncorrelated behavior is partly ensured by the use of Bootstrap Aggregation or bagging providing the randomness required to produce robust and uncorrelated trees. 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.

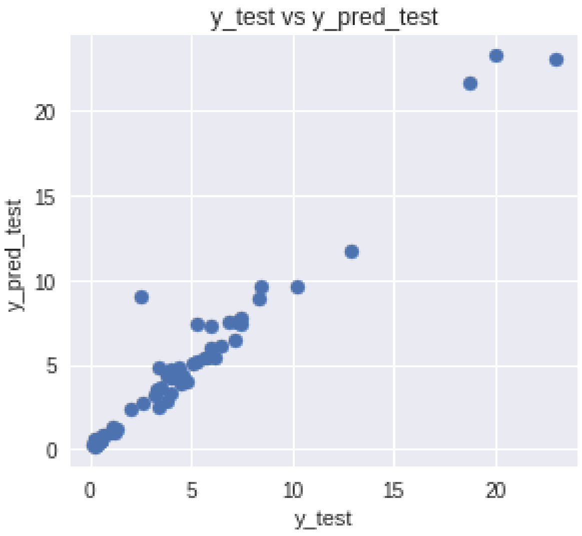




5)Gradient Boosting

Gradient Boosting is another decision tree based method that is generally described as “a method of transforming weak learners into strong learners”. This means that like a typical boosting method, observations are assigned different weights and based on certain metrics, the weights of difficult to predict observations are increased and then fed into another tree to be trained. In this case the metric is the gradient of the loss function. This model was chosen to account for non-linear relationships between the features and predicted price, by splitting the data into 100 regions.





- Interpretation of the Results

	Model	R Squared(Train)	R Squared(Test)	CV score mean(Train)
1	LinearRegression	0.88	0.86	0.84
2	Ridge	0.88	0.86	0.84
3	Lasso	0.88	0.86	0.84
4	RandomForestRegressor	0.95	0.82	0.87
5	GradientBoostingRegressor	1.00	0.94	0.87

CONCLUSION

- Key Findings and Conclusions of the Study

  Car price prediction can be a challenging task due to the high number of attributes that should be considered for the accurate prediction. The major step in the prediction process is collection and preprocessing of the data. In this project, web scraping were built to normalize, standardize and clean data to avoid unnecessary noise for machine learning algorithms.

- Limitations of this work and Scope for Future Work

As a part of future work, we aim at the variable choices over the algorithms that were used in the project. We could only explore two algorithms whereas many other algorithms which exist and might be more accurate. More specifications will be added in a system or providing more accuracy in terms of price in the system.