VISVESVARAYA TECHNOLOGICAL UNIVERSITY
"JNANA SANGAMA", BELAGAVI – 590018

Artificial Intelligence and Machine Learning (18CS71)

Assignment

Submitted in partial fulfillment of the requirements for IV Semester of

BACHELOR OF ENGINEERING IN
INFORMATION SCIENCE & ENGINEERING

Under the Guidance of

Mrs. Suchetha G

Assistant Professor

Department of Information Science & Engineering at

SAHYADRI
College of Engineering & Management
Mangaluru – 575 002

1. Introduction

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 various cities.

Understanding if a used car is worth the posted price when one sees the listings online can be difficult. Several factors, including mileage, make, model, year, etc. can influence the actual worth of a car. From the perspective of a seller, it is also a dilemma to price a used car appropriately. Based on existing data, the aim is to use machine learning algorithms to develop models for predicting used car prices.

This work "Car Price Prediction Model" focuses on building a car price prediction system using machine learning with Python. The problem statement is to build a model to predict the car prices of several used cars considering several information about the cars from a particular data set. The information can be of which car brand, a particular model, what price the car is sold and the present price of a car, the total kilometers driven, fuel type (whether it runs on Petrol, diesel or compressed natural gas), seller type (whether the seller is an individual or a dealer), transmission, whether the gear is automatic or manual and then the number of owners of that particular car, and so on.

The dataset used in the "Car Price Prediction Model" contains information about used cars listed on www.cardekho.com This data can be used for a lot of purposes such as price prediction to exemplify the use of linear regression in Machine Learning. The columns in the given dataset are as follows: Car_Name, Year, Selling_Price, Present_Price, Kms_Driven, Fuel_Type, Seller_Type, Transmission and Owner.

2. Procedure

This "Car Price Prediction Model" aims to develop a good regression model to offer accurate prediction of car price. In order to do this, we need some previous data of used cars for which we use price and some other standard attributes. Car price is considered as the dependent variable while other attributes as the independent variables.

Workflow of this particular project includes the first step as Data collection, where the car data set which has to be fed to the machine learning model is acquired. After which, the data needs to be preprocessed because one cannot feed the raw data to a machine learning algorithm. Thus, to understand the data it needs to undergo some pre-processing steps and this is called Data Preprocessing. After the data is processed, it will be split into training data and testing data. This step is called the Train Test Split. The splitter data obtained will be used to train the machine learning model. Further, testing or evaluation of the machine learning model with data is done.

Car Data → Data pre processing → Train Test split

Linear & Lasso Regression model
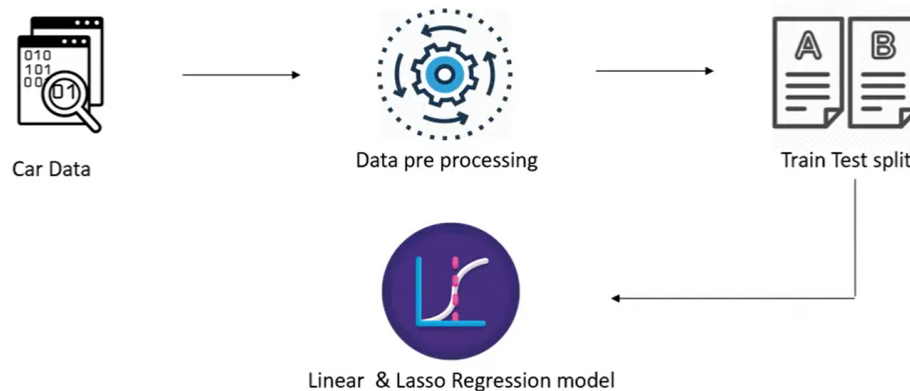
The data after the train-test split is fed into the regression model. In "Car Price Prediction Model" two regression models are going to be used. First training of the machine learning model will be done with Linear Regression and then the Lasso Regression model is used. Later, a comparison of the accuracy score or the error percentage of these models are done so as to see which model performs better to this particular data set.

3. Implementation

The implementation details of "Car Price Prediction Model" are as follows:

1. Importing the Dependencies: The required python dependencies like Pandas, matplotlib.pyplot, seaborn, sklearn, model_selection, linear_model, linear_model and Lasso are imported.

```
[43] import pandas as pd
     import matplotlib.pyplot as plt
     import seaborn as sns
     from sklearn.model_selection import train_test_split
     from sklearn.linear_model import LinearRegression
     from sklearn.linear_model import Lasso
     from sklearn import metrics
```

2. Data Collection and Processing: The loading of the data from csv file to pandas dataframe and inspecting the first 5 rows of the data frame followed by checking the number of rows and columns, getting some information about the dataset, checking the number of missing values and checking the distribution of categorical data is performed.

```
[45] # inspecting the first 5 rows of the dataframe
     car_dataset.head()
```

	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
0	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
1	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
2	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
3	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
4	swift	2014	4.60	6.87	42450	Diesel	Dealer	Manual	0

3. Encoding the Categorical Data: encoding of columns like "Fuel_Type", "Seller_Type", "Transmission" Column and so on.

```
[51] car_dataset.head()
```

	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
0	ritz	2014	3.35	5.59	27000	0	0	0	0
1	sx4	2013	4.75	9.54	43000	1	0	0	0
2	ciaz	2017	7.25	9.85	6900	0	0	0	0
3	wagon r	2011	2.85	4.15	5200	0	0	0	0
4	swift	2014	4.60	6.87	42450	1	0	0	0

4. Splitting the data and Target: Data splitting is when data is divided into two or more subsets. Typically, with a two-part split, one part is used to evaluate or test the data and the other to train the model.

```
[53] print(X)

         Year  Present_Price  Kms_Driven  Fuel_Type  Seller_Type  Transmission  \
    0    2014           5.59       27000          0            0             0
    1    2013           9.54       43000          1            0             0
    2    2017           9.85        6900          0            0             0
    3    2011           4.15        5200          0            0             0
    4    2014           6.87       42450          1            0             0
    ..    ...            ...         ...        ...          ...           ...
    296  2016          11.60       33988          1            0             0
    297  2015           5.90       60000          0            0             0
    298  2009          11.00       87934          0            0             0
    299  2017          12.50        9000          1            0             0
    300  2016           5.90        5464          0            0             0

         Owner
    0        0
    1        0
    2        0
    3        0
    4        0
    ..     ...
    296      0
    297      0
    298      0
    299      0
```

5. Splitting Training and Test data: The train-test split is used to estimate the performance of machine learning algorithms that are applicable for prediction-based Algorithms.

6. Model Training: Model training is the phase in the data science development lifecycle where practitioners try to fit the best combination of weights and bias to a machine learning algorithm to minimize a loss function over the prediction range. The two model training algorithms used here are:

- Linear Regression : Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis.

- Lasso Regression : Lasso regression is a machine learning algorithm that can be used to perform linear regression while also reducing the number of features used in the model. Lasso stands for least absolute shrinkage and selection operator.

4. Results

The results of the tests for "Car Price Prediction Model" were quantified in the following terms:

1. Linear regression :
   A linear regression model is a conditional model in which the output variable is a linear function of the input variables and of an unobservable error term that adds noise to the relationship between inputs and outputs. The output for "Car Price Prediction Model" is as follows:
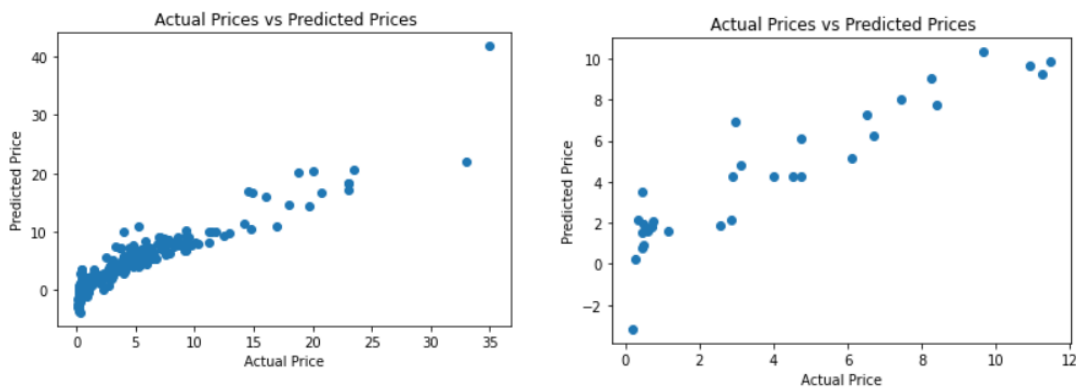
2. Lasso regression :
   Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminated from the model. The output for "Car Price Prediction Model" is as follows:
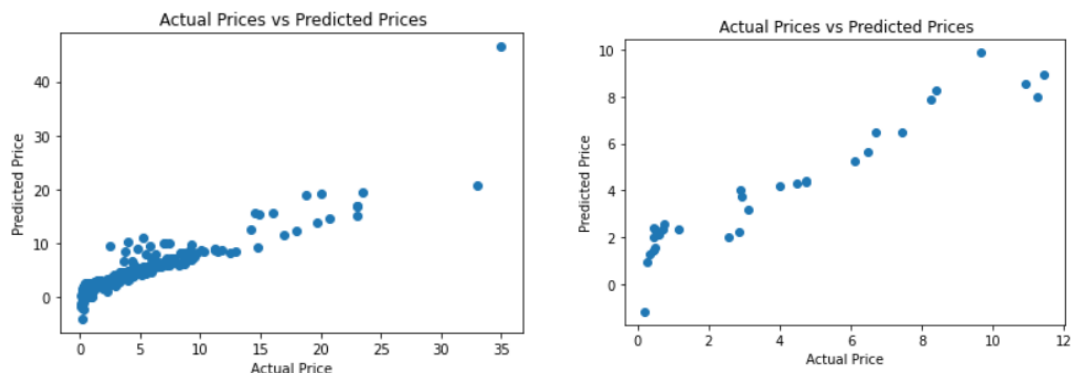
5. Conclusion

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.

India's used-car market is booming as consumers have a large variety of choices, straightforward funding, convenient digital sales channels, and a growing preference for private quality within the COVID-19 era costs of latest cars had been inflated considerably in the past few years. and so, consumers have understood that a second user automotive could be a higher value for cash instead. The pandemic has solely pushed that understanding more. So, deciding whether or not a second user automotive is well worth the denote worth once listings online will be tough. Many factors, as well as mileage, model, year, etc., will influence the particular price of an automobile. Moreover, from the attitude of a trafficker, it's additionally a perplexity to cost a second user automotive befittingly.

Due to the unprecedented number of cars being purchased and sold, used car price prediction is a topic of high interest. Because of the affordability of used cars in developing countries, people tend to purchase used cars. A primary objective of this project is to estimate used car prices by using attributes Using data mining and machine learning approaches, this project proposed a scalable framework for cars price prediction. An efficient machine learning model is built by training, testing, and evaluating two machine learning regressors.

Each experiment was performed in real-time within the Google Colab environment. In comparison to the system's integrated Jupyter notebook and Anaconda's platform, algorithms took less training time in Google Colab

6. Impact of the Assignment

In recent years, online used car trading platforms have developed rapidly, but they still face many problems. In practice, institutions and individuals differ in how they screen the characteristic variables of used car prices and predict used car prices. Under such conditions, it is easy to lead to the unsound development of the market, and it is difficult to establish a unified evaluation system, which causes great difficulties in the transaction of used cars. In terms of theory, traditional used car price evaluation methods rely too much on the subjective judgment of evaluators, which can no longer meet the needs of online transactions in the used car market. Therefore, it is necessary to establish an efficient, reasonable, fair, and accurate used car price evaluation system.