Car Price Prediction Using Machine Learning Algorithms

Suman Kumar Pal
Regd. No: 2261020145
Department of Computer Application
Siksha 'O' Anusandhan University,
Jagamara,751030
suman.pal@iter.soa.ac.in

Guided By:
Dr. Tripti Swarnkar
Professor,
Department of Computer Application

***Abstract*— As a result of incredible technological advancements and research of new technical expertise and huge economical growth of our country, people started to buy cars more than other vehicles. Therefore, there arises an enormous demand for cars, as the demand increases for new cars the used car market also booms alike. But the used car market is highly manipulated by few numbers of people who govern the rates of the used cars and also online selling websites designate the values for the used cars. This paper tries to study and investigate the trends in used car prices and predicts the price of used cars with the help of supervised machine learning algorithms. And to suggest which machine learning algorithm performs well among the selected methods for predicting the cars price. There has been related work done with machine learning algorithms like linear regression, multiple regression, random forest and so on. We wanted to study which algorithm predicts the car price more reliably and accurately. So that this solution will be helpful for first time used car buyers and also for sellers for determining the selling cost of the car.**

Keywords — Keywords: Supervised Machine Learning Algorithms, Linear Regression, Multiple Regression, Random Forest;

I. Introduction

In India the automobile market is a biggest business for international and Indian automobile companies. As the boom and demand for automobiles increase there is also a big market opening for used cars. The used car market is being manipulated and controlled by some of the online advertisement websites like olx and quickr, but customers who want to buy a used car is easily being manipulated and cheated to a higher price which the car isn't worth buying for. I would like to propose a solution for this problem by using the help of artificial intelligence and machine learning by using some supervised learning machine learning techniques and algorithms to predict the used car prices based on some parameters. And I want to investigate and compare the accuracy which different algorithms produce on testing and predicting with the used car data. During 2019-20 the entire automobile production in India was 26,353,293, But in 2020-21 the automobile production in India was 22,652,108. We can see that there is a huge decline in automobile industry, people are preferring more on used and second-hand vehicles than new vehicles. Therefore, the system of used cars must be standardized and a clear pricing system needs to be implemented. This paper suggests few machine techniques which can be used to predict the prices of used cars with historical used car prices data and considering a mean value from the list of prices for a specific car and assigning it as the predicted price for the given features and parameters. There has been many related work-done regarding this topic and field but only very few or one or two authors have done for Indian dataset, Thus I wanted to find a solution for this problem and find prediction method to give the prices for used cars in a correct method. The data for this car price prediction experiment is taken from various sources like Kaggle, web-scrapping and open-source data websites which provide free data. A car price prediction has been a high-interest research area, as it requires noticeable effort and knowledge of the field expert. Considerable number of distinct attributes are examined for the reliable and accurate prediction. As the demand for cars increase the demand for second hand and used cars also increases so due to this high demand, we need to build a AI solution for solving this demand in a customer friendly way. The customers are getting cheated and tricked for a higher price for a less worth used car if the customer wants to buy it from a dealer who sells used cars. The dealer tries to sell a damaged or repaired car for high price to customers who don't know much about buying cars and stuff. The customer who doesn't know about the technical specifications and other prices of spare parts and how to deduct the price will easily be cheated with high price. I wanted to solve this type of problem where the customer has to know exact price the car is worth for. Using machine learning it is possible to predict the correct and worthy price for a given used car based on previous data from various sellers and buyer.

This can be done by training the model using used cars dataset which has several features and parameters such as year of manufacturing, model year, number of cylinders, number of kms/miles driven, diesel or petrol, automatic or manual or other type of transmission, the gearing system of the cars, the number of owners of the car etc., like this there are many features from which the cars price can be predicted. And also we can add if there is any damage or is it flood affected or accidental damaged car these factors can also be considered for predicting the correct and exact price of the car.

II. MACHINE LEARNING

Machine Learning is one of efficient technology which is based on two terms namely testing and training i.e. system take training directly from data and experience and based on this training test should be applied on different type of need as per the algorithm required.

There are three type of machine learning algorithms:

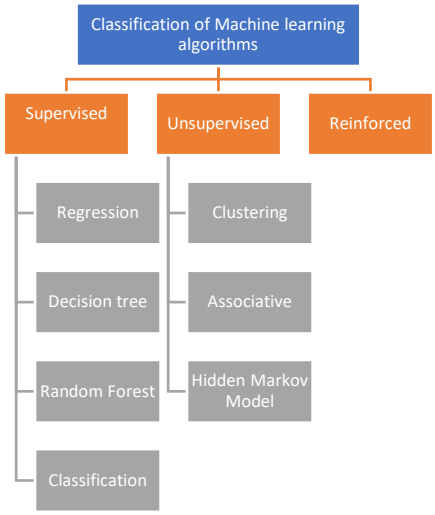

Fig.1 Classification of machine learning

A. Supervised Learning

Supervised learning can be defined as learning with the proper guide or you can say that learning in the present of teacher. we have a training dataset which act as the teacher for prediction on the given dataset that is for testing a data there are always a training dataset. Supervised learning is based on "train me" concept. Supervised learning have following processes:

- Classification
- Random Forest
- Decision tree
- Regression

To recognize patterns and measures probability of uninterruptable outcomes, is phenomenon of regression. System have ability to identify numbers, their values and grouping sense of numbers which means width and height, etc. There are following supervised machine learning algorithms:

- Linear Regression
- Logistical Regression
- Support Vector Machines (SVM)
- Neural Networks
- Random Forest
- Gradient Boosted Trees
- Decision Trees
- Naive Bayes

B. Unsupervised Learning

Unsupervised learning can be defined as the learning without a guidance which in Unsupervised learning there are no teacher are guiding. In Unsupervised learning when a dataset is given it automatically work on the dataset and find the pattern and relationship between them and according to the created relationships, when new data is given it classify them and store in one of them relation. Unsupervised learning is based on "self-sufficient " concept.

For example: consider a dataset containing information about customer purchases in a supermarket, including items like bread, milk, eggs, vegetables, and snacks. When applying unsupervised learning, clustering algorithms can be used to group similar purchasing patterns together.

For instance, using clustering techniques like k-means, the algorithm may identify three clusters: one for customers who primarily buy fresh produce and vegetables, another for those who focus on dairy and bakery items, and a third for customers who frequently purchase snacks and non-perishable goods. Now, if a new customer's purchasing history is provided, the algorithm can automatically assign them to one of these clusters based on their buying behaviour. This can be valuable for targeted marketing, personalized recommendations, or inventory management.

In this scenario, the algorithm learns patterns and relationships in the data without explicit labels, helping to uncover underlying structures and groupings in the customers' purchasing habits. Un-supervised algorithms have following process:

- Dimensionality
- Clustering

There are following unsupervised machine learning algorithms:

- k-means clustering
- PCA

C. Reinforcement

Reinforced learning is the agent ability to interact with the environment and find out the outcome. It is based on "hit and trial" concept. In reinforced learning each agent is awarded with positive and negative points and on the basis of positive points reinforced learning give the dataset output that is on the basis of positive awards it trained and on the basis of this training perform the testing on datasets

III. METHODOLOGY OF SYSTEM

A. Implementation

The implementation was divided into five parts titled Data Set, Data Cleaning and Normalization, Machine Learning Algorithms, Measurements, and Inference.

The entire implementation was written in Python3 in the PyCharm ide. The libraries utilized are pandas, sklearn (sci-kit learn), NumPy, matplotlib, and seaborn.

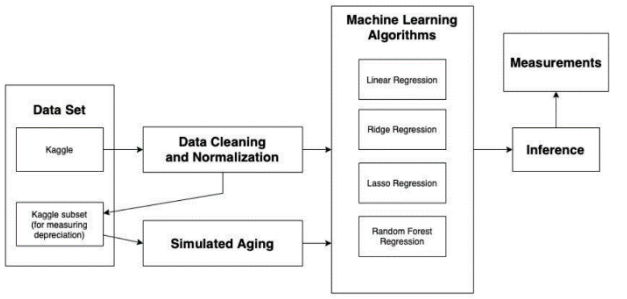

Fig.2 Architecture of Prediction System

First step for predication system is data collection(9 features & 301 samples) and deciding about the training and testing dataset. In this project we have used 90% training dataset and 10% dataset used as testing dataset the system.

B. Attribute Selection

Attribute of dataset are property of dataset which are used for system and for car price, attributes are like car name, year, selling price, present price and many more shown in TABLE.1 for predication system.

	A	B	C	D	E	F	G	H	I
1	Car_Name	Year	Selling_Pri	Present_P	Kms_Drive	Fuel_Type	Seller_Typ	Transmiss	Owner
2	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
3	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
4	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
5	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
6	swift	2014	4.6	6.87	42450	Diesel	Dealer	Manual	0
7	vitara brez	2018	9.25	9.83	2071	Diesel	Dealer	Manual	0
8	ciaz	2015	6.75	8.12	18796	Petrol	Dealer	Manual	0
9	s cross	2015	6.5	8.61	33429	Diesel	Dealer	Manual	0
10	ciaz	2016	8.75	8.89	20273	Diesel	Dealer	Manual	0
11	ciaz	2015	7.45	8.92	42367	Diesel	Dealer	Manual	0
12	alto 800	2017	2.85	3.6	2135	Petrol	Dealer	Manual	0
13	ciaz	2015	6.85	10.38	51000	Diesel	Dealer	Manual	0
14	ciaz	2015	7.5	9.94	15000	Petrol	Dealer	Automatic	0
15	ertiga	2015	6.1	7.71	26000	Petrol	Dealer	Manual	0

Fig.3 Dataset

TABLE.1 Attributes of the Dataset

S. No.	Attribute	Description	Type
1	Car Name	Name of various types of cars	
2	Year	The production year of the vehicle	
3	Selling Price	The price at which the car will be sold	Numerical
4	Present Price	Varies by model, age, condition; research for accurate market value.	Numerical
5	Fuel Type	Types of fuel that consume(Petrol, Diesel, CNG)	Nominal
6	Seller Type	Dealer, Individual	Nominal
7	Transmission	Manual, Automatic	Nominal
8	Owner	Total nos of owner	Numerical

C. Data Cleaning

The first step in cleaning the dataset provided from Kaggle was to identify variables which will not be useful for training the models. This includes features which are not correlated with price, have too many discrete values to draw inferences from, or have too many missing values. The features that were identified to be dropped from the dataset were: Car_Name, Year and Owner. The next step is identifying and removing outliers for the remaining features. Keeping in mind the distribution of the data and the negative effect of removing too many values, appropriate minimum and maximum values were set for each feature to remove rows in the dataset which were extreme in any feature category.

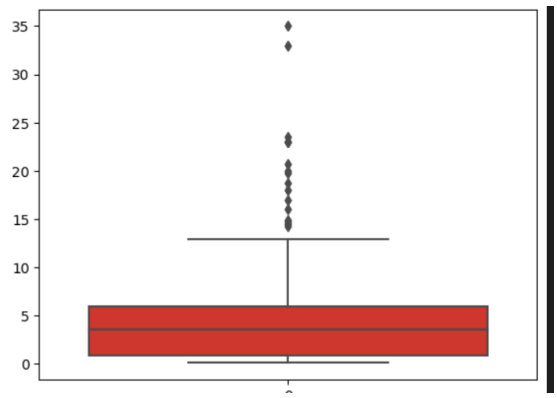

Fig.4 box-plot

	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
64	fortuner	2017	33.0	36.23	6000	Diesel	Dealer	Automatic	0
86	land cruiser	2010	35.0	92.60	78000	Diesel	Dealer	Manual	0

Fig. 5 Samples containing Outlier

	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
...	...	...	...	...	...	...	...	...	...
296	city	2016	9.50	11.60	33988	Diesel	Dealer	Manual	0
297	brio	2015	4.00	5.90	60000	Petrol	Dealer	Manual	0
298	city	2009	3.35	11.00	87934	Petrol	Dealer	Manual	0
299	city	2017	11.50	12.50	9000	Diesel	Dealer	Manual	0
300	brio	2016	5.30	5.90	5464	Petrol	Dealer	Manual	0

299 rows × 9 columns

Fig.6 Removing outlier from the dataset

D. Preprocessing of data

Preprocessing needed for achieving prestigious result from the machine learning algorithms. For example Random forest algorithm does not support null values dataset and for this we have to manage null values from original raw data.

For our project we have to convert some categorized value by dummy value means in the form of “0”and “1” by using following code:

```
#encoding the categorical data

# encoding "Fuel_Type" Column
car_dataset.replace({'Fuel_Type':{'Petrol':0,'Diesel':1,'CNG':2}},inplace=True)

# encoding "Seller_Type" Column
car_dataset.replace({'Seller_Type':{'Dealer':0,'Individual':1}},inplace=True)

# encoding "Transmission" Column
car_dataset.replace({'Transmission':{'Manual':0,'Automatic':1}},inplace=True)
```

Fig.7 change the categorical data into numerical

E. Histogram of attributes

Histogram of attributes shows the range of dataset attributes and code which is used to create it. dataset.hist() .

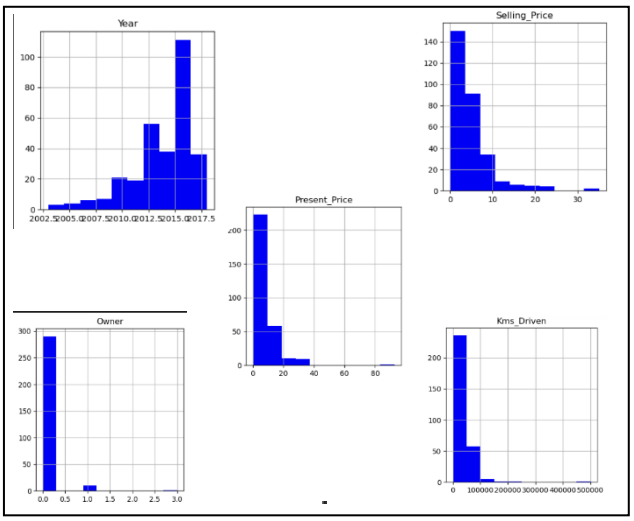

Fig.8 Histogram of attributes

IV. MACHINE LEARNING ALGORITHMS

The data, after being cleaned and normalized, is split into training and test data using a randomized 90-10 split. This is to ensure that the data used for testing does not contain any of the data used for training. Thus 10% of the data is reserved for testing purposes. The training dataset was used to train the four price prediction ML models chosen: Multiple Linear Regression, Lasso Regression, Ridge Regression, and Random Forest Regression. All machine learning algorithms used in this

report were imported from the sklearn library. Some models were provided input parameters to implement.

A. Linear regression

It is the supervised learning technique. In simple linear regression, there is one independent variable predicting a dependent variable. The equation for a simple linear regression model can be represented as:

$$E(Y|x) = \beta 0 + \beta 1x$$

Y : Dependent Variable (Selling Price)

X : Independent variable (a feature of the car, e.g. Present price, Kms_drive

β0 : Intercept (the value of Y when X is 0)

β1 : Slope (the change in Y for a unit change in X)

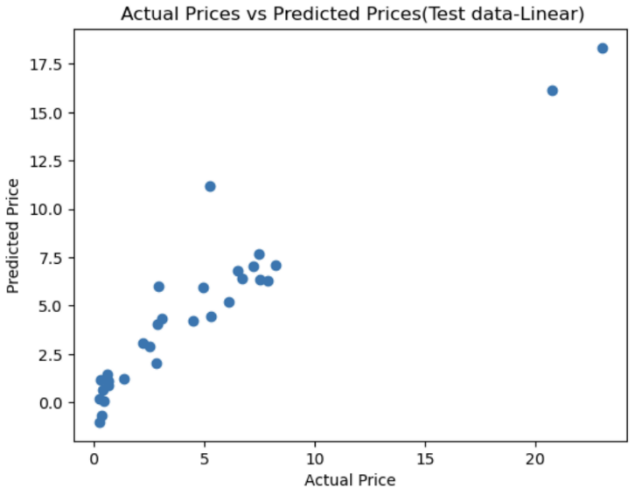

Fig.9 Difference between actual and predicted point

It gives a relation equation to predict a dependent variable value “y” based on a independent variable value “x”. so it is concluded that linear regression technique give the linear relationship between x(input) and y(output).

B. Lasso Regression

Lasso regression, short for Least Absolute Shrinkage and Selection Operator, is a type of linear regression that incorporates regularization to prevent overfitting and feature selection. Lasso regression adds a penalty term to the standard linear regression objective function, encouraging the model to use fewer features by driving the coefficients of some features to exactly zero. This makes it particularly useful when dealing with datasets that have a large number of features, as it helps in selecting the most relevant ones.

Cost fun:

$$J(\theta)=1m\sum i=1m[-y(i)\log(h\theta(x(i)))-(1-y(i))\log(1-h\theta(x(i)))]$$

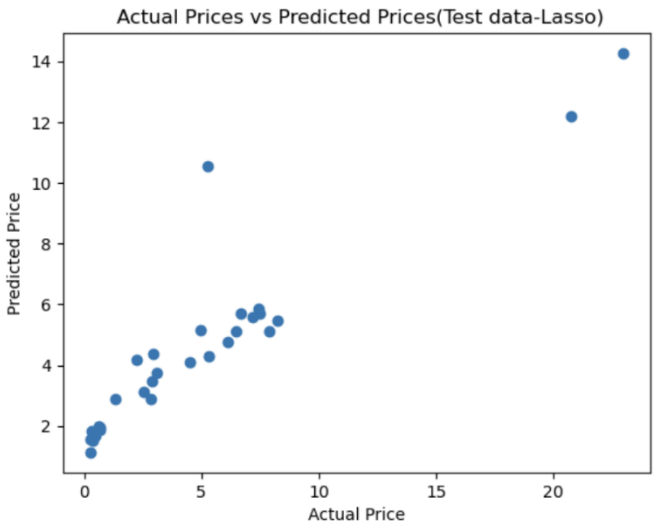

Fig.10 Lasso Regression

C. Ridge Regression

Ridge Regression is a linear regression technique that is used to address the issue of multicollinearity in the dataset. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, making it difficult to identify the individual effect of each variable on the dependent variable.

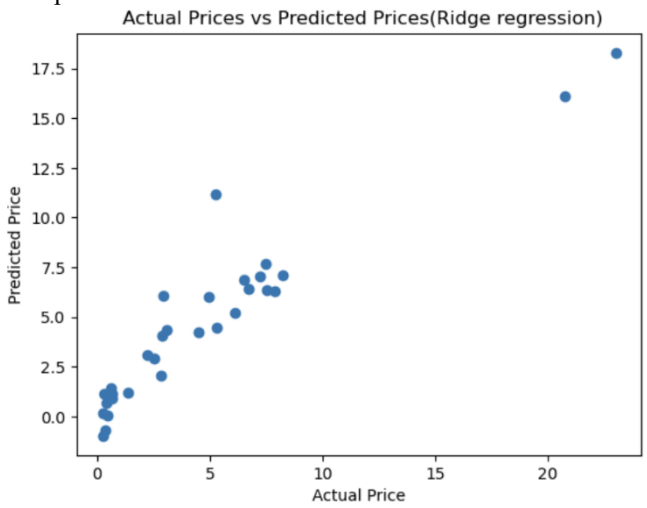

Fig.11 Ridge Regression

In the context of car price prediction, let's assume you have a dataset with various features (independent variables) such as engine size, horsepower, fuel efficiency, brand, and so on, and you want to predict the price of a car (dependent variable)

D. Decision tree

Decision trees are a popular machine learning algorithm used for both classification and regression tasks. In the context of car price prediction, we'll focus on using a decision tree for

regression. The goal is to build a model that can predict the price of a car based on various features.

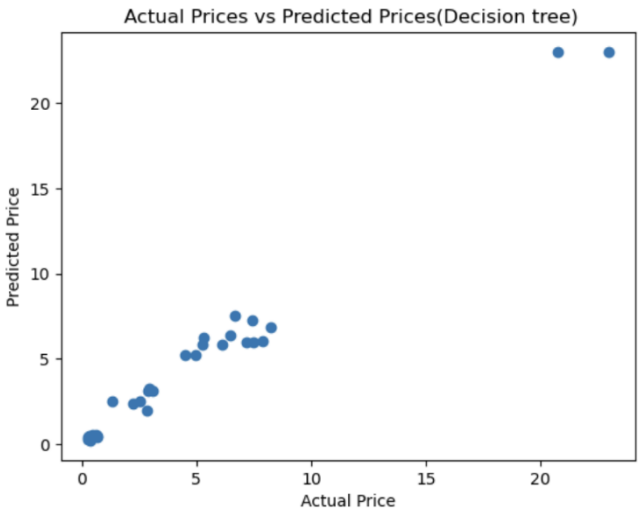

Fig.12 Decision tree

For the tree construction we use entropy of the data attributes and on the basis of attribute root and other nodes are drawn.

Entropy= -∑ Pij log Pij

E. Random Forest Regression

Random Forest Regression for car price prediction utilizes an ensemble of decision trees to enhance accuracy. Each tree independently predicts car prices, and the final prediction is an average of all trees' outputs. This method solves overfitting and increases robustness.

Features like current price, model, year, owner type and fuel type influence individual tree decisions. During training, each tree learns unique patterns, contributing to diverse predictions. The model's strength lies in combining these diverse insights, offering a reliable and accurate car price estimation. This approach handles non-linearity effectively, making it well-suited for capturing complex relationships within the dataset and improving overall prediction performance.

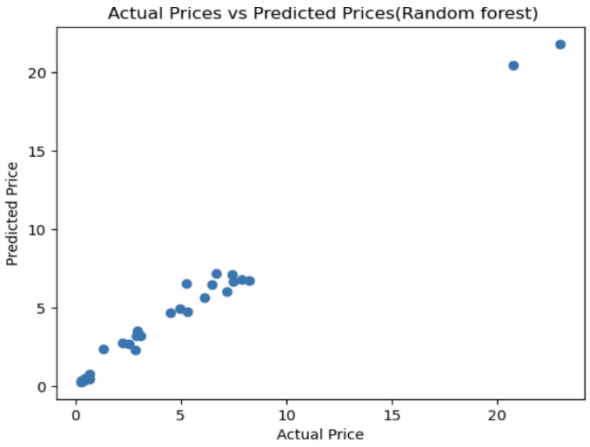

Fig.13 Random Forest

V. Result Analysis

A. About Jupyter Notebook

Jupyter Notebook serves as an effective simulation tool, offering a comfortable environment for Python programming projects. Its versatility lies in the seamless integration of rich text elements and code, encompassing figures, equations, links, and more. This amalgamation of text and code makes Jupyter Notebooks an ideal platform to consolidate analysis descriptions alongside real-time data analysis execution and result representation.

As an open-source, web-based tool, Jupyter Notebook facilitates interactive creation of graphics, maps, plots, visualizations, and narrative text. Its user-friendly interface enables users to seamlessly weave together documentation and executable code, providing a dynamic and interactive environment for data analysis projects.

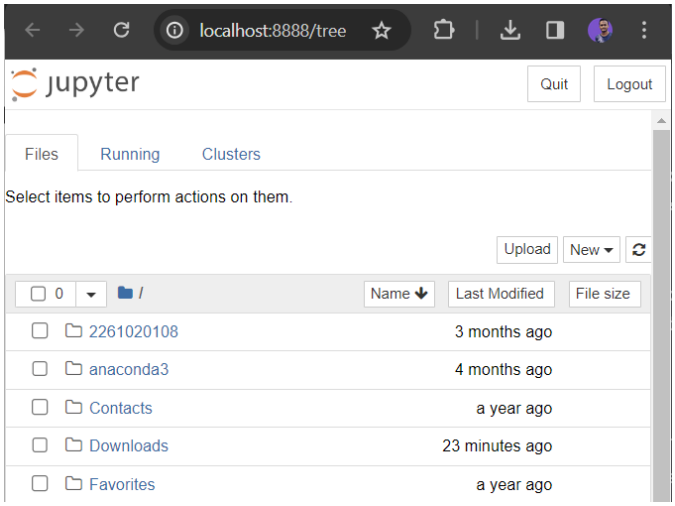

Fig.14 Jupyter Notebook

B. Results

R-squared (R^2) is a metric commonly used to assess the accuracy of regression models, including Random Forest Regression. It quantifies the proportion of variance in the dependent variable (car selling prices in this case) that the model explains. The R^2 value ranges from 0 to 1, with higher values indicating better fit.

If the Random Forest Regression model has an R-squared value close to 1, it suggests that a significant proportion of the variability in car prices is accounted for by the model, indicating high accuracy. Conversely, an R-squared value closer to 0 implies lower accuracy, meaning the model doesn't explain much of the variance in the target variable. Regularly validating R^2 alongside other metrics ensures a comprehensive evaluation of the model's performance in predicting car prices.

After performing the machine learning approach for testing and training we find that performance of the Random forest regression is much efficient as compare to other algorithms. It is conclude that Random forest is best among them with 0.982 performance and the comparison is shown in TABLE.2.

TABLE.2 Performance comparison

Algorithm	R^2 _Score
Linear Regression	0.876212
Lasso Regression	0.725747
Ridge Regression	0.875396
Random Forest	0.982096
Decision Tree	0.977032

VI. CONCLUSION AND FUTURE SCOPE

A. *Conclusion*

The first process was to determine which of the models and parameters gives the best overall accuracy/performance in making price predictions for used cars. The optimal parameters were determined in the process of implementing the models, and thus each model was implemented with the parameters that yielded the best performance by trial and error. The results show that out of the five models tested, Random Forest Regression provided the highest accuracy in all of the metrics used and highest overall accuracy.

The second process was to determine which of the models can most accurately assess the depreciation of a car over time. All of the models approximated geometric appreciation, meaning that a constant percentage of value is lost every year independent of the age of the vehicle.

The third process is to determine which model demonstrates the best potential for development of a consumer tool for evaluating used cars or a particular subset of used cars. The results show that Random Forest Regression performed the best on all performance metrics and for all price percentile subsets of used cars. It was also much better able to approximate the depreciation.

B. *Future works*

Applying the Method to Other ML Models, This work compared the performance of four ML Regression algorithms. A way to expand this work in the future is to apply the same method for comparing these algorithms to others that are suited to regression problems. Some example algorithms are Light Gradient Boosted Machine (LGBM), Kth Nearest Neighbour Regression (KNN), Decision Tree Regression (DTR), and Artificial Neural Networks (ANN). The problem of price prediction deals with continuous variables which makes it suited to regression algorithms, but by creating discrete intervals for the continuous variables such as price, other algorithms could be applied.

Adding Additional Features Related to the Year, A potential improvement to the predictive power of all ML models, if they are able to take advantage of the information, is to add more correlated features. There are some features which are not related to the attributes of the car, such as the price of fuel. A car that uses more fuel will be worth less when fuel costs more. Other such features could include the economic conditions, or changes in the climate.

References

[1] **Dataset :** https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho

[2] https://ieeexplore.ieee.org/document/9696839

[3] https://thecleverprogrammer.com/2021/08/04/car-price-prediction-with-machine-learning/

[4] https://github.com/topics/car-price-prediction-with-machine-learning

[5] https://towardsdatascience.com/predicting-car-price-using-machine-learning-8d2df3898f16

[6] https://www.researchgate.net/publication/335799148_Car_Price_Prediction_Using_Machine_Learning