Predicting Car Selling Price (Machine Learning Approach)

1.Introduction:

Determining the best price to sell used automobiles for in the ever-changing automotive market is an essential task. The objective of this project is to create a machine learning model that can correctly forecast the selling price of cars based on a variety of factors. The dataset utilized in this study includes details about the specs of the cars, including the year of purchase, the price at the time of use, the fuel type, the transmission, the seller type, and more.

2. Methodologies:

1.Understanding the Dataset:

The data is obtained from Cardekho which consists of 302 rows of data with columns carname,Year,Sellingprice,Present price,kilometers driven,Fuel type,Seller type,Transmission ,Owner with no null values.

2.Data Preprocessing:

Eliminating Irrelevant Features:

The "Car_Name" column is removed from the code because it has no bearing on the predictive modeling. The information is used for this.declaration of drop("Car_Name", axis=1, inplace=True).

Encoding Categorical Variables:

Categorical variables are encoded using the scikit-learn LabelEncoder. The encoded categorical variables are "Fuel_Type," "Transmission," and "Seller_Type." The category labels are changed into numerical representations in this step, which are better suited for machine learning.

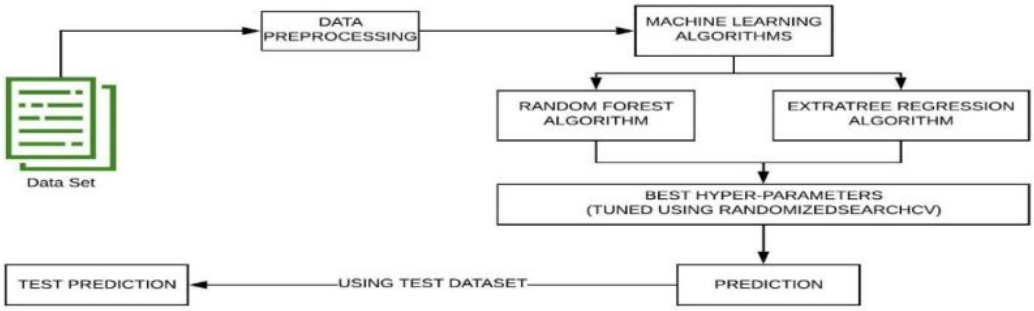
Feature Scaling:

Using the MinMaxScaler and StandardScaler from scikit-learn, two numerical features, "Present_Price" and "Kms_Driven," go through feature scaling. By bringing the values of these characteristics into a narrow range, feature scaling stops any one feature from predominating over others during model training.

3.Data Training and Modelling:

Using the train_test_split function from scikit-learn, the data is divided into training and testing sets. The testing set is used to assess the machine learning model's performance, whereas the training set is utilized to fit the models.
Additional data pretreatment stages, like handling missing values, feature engineering, and outlier detection, may be required depending on the unique dataset and the specifications of the prediction model. In order to create accurate and reliable machine learning models, data preparation is an essential step that frequently necessitates careful attention based on the features of the available data.
30% of data used to testing and 70% of data used for training
Algorthims we used to build the model Linear Regression,Random Forest,XgBoost

4.Proposed Model :



The proposed model is an application of the two machine learning algorithms i.e. Random Forest Algorithm and Linear Regression algorithm. In this model first, the dataset is loaded for further exploration. In this specific model, We used a Dataset available at Cardekho . we start training the model for distributed dataset into two

1. Training Dataset
2. Test Dataset.

Applied the two machine Learning algorithms i.e. Random Forest Algorithm and Xgboost Algorithm . Once the model predicts a result, We test the prediction using test dataset created using the scikit-Learn library and calculate its accuracy

Using evaluation matrics we found that XgBoost algorithm surpasses others

XgBoost r2 score 0.96 and mean absolute 0.618 and mean squared error 1.07 and root mean squared error 1.03

3.Project Overview:

Using the pandas, seaborn, and matplotlib packages, the project starts with data loading and early exploration. In order to better comprehend the properties of the data, visualizations are used to get insights into the distribution of important features. The 'Car_Name' field has been removed because it has no bearing on the predictive modeling.

Using LabelEncoder from scikit-learn, the categorical variables "Fuel_Type," "Transmission," and "Seller_Type" are encoded into numerical form appropriate for machine learning techniques. Additionally, numerical aspects like "Present_Price" and "Kms_Driven" are scaled with the help of MinMaxScaler and StandardScaler to keep the values within a particular range and avoid the dominance of one feature over another.

Using train_test_split from scikit-learn, the dataset is then divided into training and testing sets. Three distinct regression models—Linear Regression, RandomForestRegressor, and XgBoost—are fitted using the training set. To maximize performance, the models are carefully set up with particular hyperparameters, such as the number of estimators, the maximum depth, and the minimum samples leaf.

To determine each model's propensity for prediction, evaluation is essential. The accuracy of the predictions is measured using metrics like R-squared, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). This makes it possible to compare the models and finds the one that is best for forecasting car selling prices.

The XgBoostmodel surpasses the others, showing the best R-squared value and the lowest MAE, MSE, and RMSE, according to the data. This shows that the most reliable model for this particular prediction task is the XgBoost. The model helps automobile sellers and buyers

by offering insightful information on the critical elements influencing the selling price of the vehicle.

Users can input individual car parameters to get the associated anticipated selling price, illustrating the practical application of the trained XgBoost model. With the addition of interactive features, the project becomes more approachable and applicable to real-world situations.

4.Conclusion :

This Project demonstrates how machine learning can be used to estimate car selling prices based on a variety of factors. The performance of the model is thoroughly analyzed through the use of several regression models and evaluation measures, allowing players in the automobile sector to optimize their pricing strategies and make well-informed decisions. The project's utility is further improved by the interactive prediction element, making it a useful resource for both buyers and sellers of cars.

References:

N.Donges –“ A COMPLETE GUIDE TO THE RANDOM FOREST ALGORITHM”
https://builtin.com/data-science/random-forest-algorithm

A. Dey – “Data Pre-processing for Machine Learning”
https://medium.com/datadriveninvestor/data-preprocessing-for-machine-learning-188e9eef1d2c

T. Yiu –“ Understanding Random Forest How the Algorithm Works and Why It Is So Effective"
https://towardsdatascience.com/understanding-random-forest-58381e0602d2

This project is developed by Phoenix softech limtd. This project is developed by a team of interns guided by lead developer Karthika K. The interns have majorly contributed to the project.

Company:www.phoenixsoftech.in
Mentor : Karthika K,
Lead Developer,
Phoenix Softech Limited.
Interns:
Hariharan R K
Vishnu Kumar M J
Vijay Purushoth M
Saffryn Timothy S
Dhanaseelan V