International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022

Forecasting Vehicle Prices using Machine Learning Techniques

Ms. R. Uma
Lecturer, Information Technology
PSG Polytechnic College, Tamil Nadu, India

J. Kamal
Information Technology
PSG Polytechnic College, Tamil Nadu, India

G. Sri Siva Thandavan
Information Technology
PSG Polytechnic College, Tamil Nadu, India

S. Raghul
Information Technology
PSG Polytechnic College, Tamil Nadu, India

Abstract: **A car price prediction has been a high-interest research area as it requires noticeable effort and knowledge of the field expert. Considerable numbers of distinct attributes are examined for the reliable and accurate prediction. To build a model for predicting the price of used cars, three machine learning techniques (Artificial Neural Network, Support Vector Machine and Random Forest) are applied. However, the mentioned techniques are applied to work as an ensemble. The data used for the prediction was collected from the web portal autopijaca.ba using web scraper that was written in PHP programming language. Respective performances of different algorithms were then compared to find the one that best suits the available data set. The final prediction model was integrated into Java application. Furthermore, the model was evaluated using test data and the accuracy of 87.38% was obtained.**

Keywords: Random Forest, XGBoost, LightGBM, CatBoost, Extra Trees

I. INTRODUCTION

From a long time, due to the fact that being a non-stop paradigm of transactions, commodities have been into lifestyles. Earlier those transactions hadbeen within the form of barter gadget which later was translated into a economic machine. And with consideration into these all modifications that were introduced about the pattern of re-promoting objects changed into effects as well. There are approaches wherein the re-selling of the object is finished.

One is offline and the opposite being online. In offline transactions, there is a mediator found in among who's very at risk of being corrupt and make overly profitable transactions.

The second option is online in which there may be a certain platform which shall be the person find the price he might get if he is going for promoting. Kilometers traveled – The number of kilometers travelled with the aid of an automobile has a massive function to play while putting the vehicle up on the market. The extra the automobile has travelled, the older it's miles. Fiscal energy – It is the electricity output of the automobile. More output yields higher price out of an automobile.

Year of registration – It is the year when the vehicle was registered with the Road Transport Authority. The more modern the automobile is, the better value it's going to yield.By every passing year, the value of the vehicle will depreciate. Fuel Type – There are two types of fuels present in the data set which is going to be used in this project namely Petrol and Diesel. A system that can develop a self-learning/machine-learning is needed to predict the price of the used cars. To build a supervised machine learning model for forecasting value of a vehicle based on multiple attributes is the main objective of this project which is going to be a real-time task.

Car price prediction is extremely an interesting and a well-known problem. 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. This adds additional importance to the problem of the auto price prediction. Accurate car price prediction entails expert knowledge because rate usually depends on many different factors. Typically, most significant ones are brand, model, age, horsepower, mileage as well as economics of demand and supply factors.

The gasoline used in the vehicle has an effect on the price of an automobile due to frequent increase in the rate of fuels viz. Diesel, Petrol & CNG. Different technical, agronomical, figurative features like outdoors color, door range, sort of transmission, dimensions, passenger safety like airbags, air conditioning, interiors like upholstery, navigation gadget, infotainment gadget, and cruise manipulate and so forth may also have an effect on automobile costs.

II. PROPOSED SYSTEM

There are majorly two features provided in this project namely 1. Re-sale platform: A centralized platform for car resale that will predict prices. 2. Feature selection: Feature-based search and prediction. The process starts by collecting the dataset. The next step is to do Data Preprocessing which includes Data cleaning, Data reduction, Data Transformation. Then, using various machine learning algorithms we will predict the price. The algorithms involve Linear Regression, Ridge Regression and Lasso Regression. The best model which predicts the most accurate price is selected. After selection of the best model the predicted price is displayed to the user according to user’s inputs. User can give input through website to be used for car price prediction. This process compares the accuracy score of Decision Tree, Logistic Regression, Random Forest, Gradient Boosting Algorithms and Naive Bayes algorithms etc,. for predicting the used car price. The block diagram and work flow of the proposed system are shown in Fig 2.1 and 2.2.

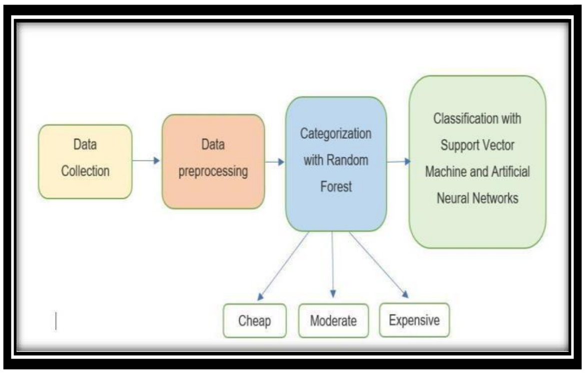

Data Collection
Data preprocessing
Categorization with Random Forest
Classification with Support Vector Machine and Artificial Neural Networks
Cheap
Moderate
Expensive

Fig 2.1 Block Diagram

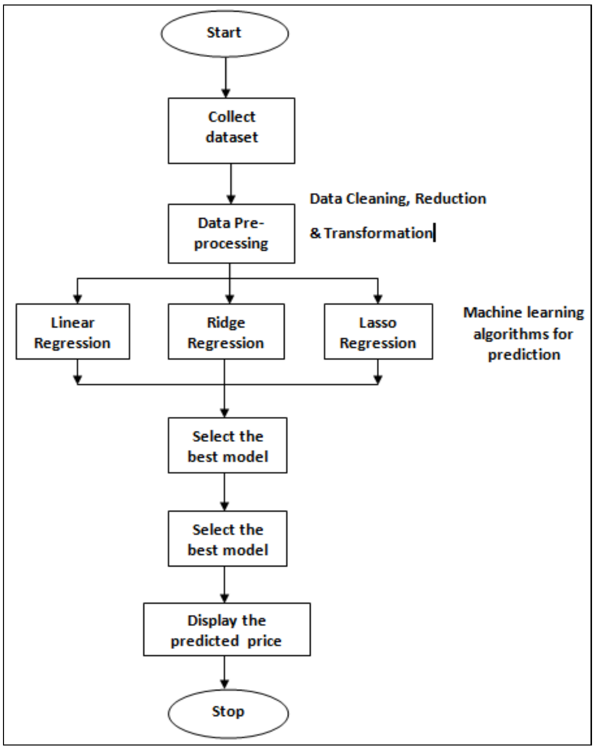


Fig 2.2 Work flow diagram

III. IMPLEMENTATION

In this project various machine learning algorithms such as Random forest, XGBoost, LightGBM, CatBoost and Extra Trees are implemented over the dataset containing the used car parameters and the accuracy is measured.

Random forest

Random forest algorithmcreates decision trees on data samples during the training and then gets the prediction from each of them. Finally it selects the best solution by means of voting. It takes their majority vote for classification and average in case of regression.This ensemble method is better than a single decision tree because it reduces the over-fitting by averaging the result. That is, it combines predictions from multiple machine learning algorithms to make a more accurate prediction.

Working of Random Forest Algorithm

Step 1: Random samples are selected from the given dataset.
Step 2: Individual decision trees are constructed for each sample and prediction result is obtained from each decision tree.
Step 3: Voting will be performed for every predicted result.
Step4: Most voted prediction result is selected as the final prediction result.

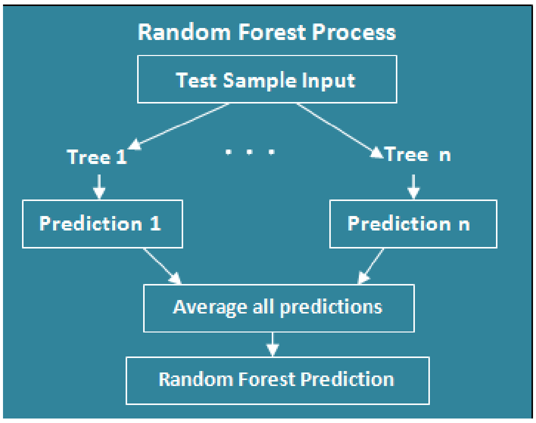


Fig. 3.1 Random Forest process

XGBoost

Extreme Gradient Boosting creates decisiontrees in sequential form as shown in Fig. 3.2. All the independent variables are assigned with weights and these weights are fed into the decision tree which predicts results. The wrongly predicted weight of the variables are then fed to the second decision tree. Theses individual predictors ensemble to provide a more precise model.

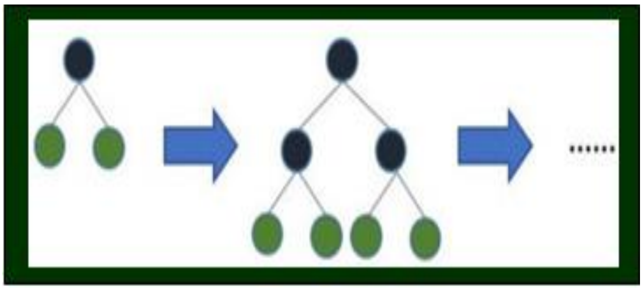


Fig. 3.2 Level-wise tree growth

LightGBM (Light Gradient Boosting Machine)

Light Gradient Boosting Machine algorithm is a gradient boosting framework based on decision trees as shown in Fig. 3.3. It increases the efficiency of the model and at the same time reduces the memory usage. It is much better than XGBoost when dealing with large datasets. The main objective of this algorithm is to get good accuracy of results.

It uses two novel techniques:Gradient-based One Side Sampling(GOSS) and Exclusive Feature Bundling (EFB) which fulfills the limitations of histogram-based algorithm that is primarily used in all GBDT (Gradient Boosting Decision Tree) frameworks.

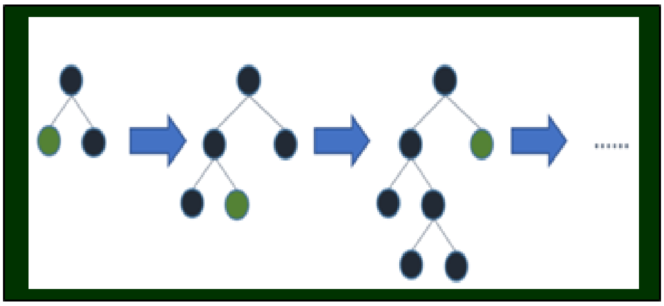


Fig. 3.3 Leaf-wise tree growth

The techniques of GOSS and EFB form the characteristics of LightGBM Algorithm. They comprise together to make the model work efficiently and provide it a cutting edge over other GBDT frameworks.

CatBoost

CatBoost is based on gradient boosted decision trees. During training, a set of decision trees is built consecutively. Each successive tree is built with reduced loss compared to the previous trees.

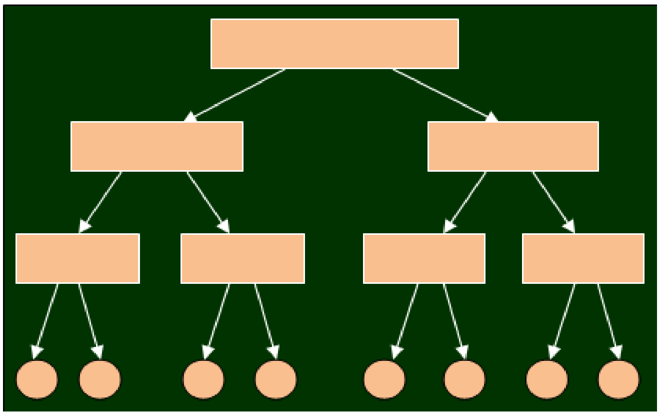


Fig. 3.4 Oblivious trees in CatBoost

This algorithm grows oblivious trees as shown in Fig. 3.4. Which means that the trees are grown based on the rule that all the nodes at the same level, test the same predictor with the same condition, and hence the index of a leaf can be calculated with bitwise operations. This oblivious tree method allows simple fitting scheme and efficiency on CPUs to find an optimal solution and avoid over fitting.

Extra Trees

The Extra Trees algorithm aggregates the results of multiple de-correlated decision trees collected in a forest to output it's classification result as shown in Fig. 3.5. Here each decision tree is built from the original training sample. Predictions are made by averaging the prediction of the decision trees in the case of regression or using majority voting in the case of classification.

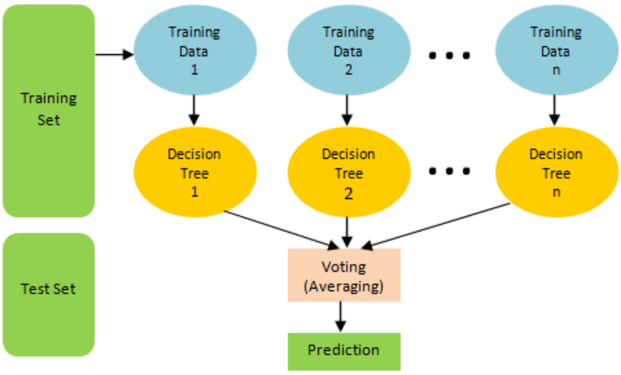
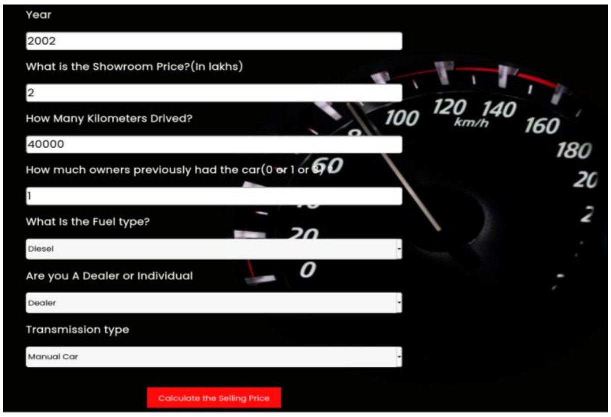


Fig. 3.5 Prediction process in Extra Trees

This algorithm randomly sample the features at each split point of a decision tree. It selects a split point at random. The three main hyper parameters to tune are: the number of decision trees in the ensemble, the number of input features to randomly select and consider for each split point and the minimum number of samples required in a node to create a new split point.

IV. EXPERIMENTAL EVALUATION

The user interface to provide input data is as shown in Fig. 4.1. It includes the year, price of the car, distance travelled, number of previous owners, fuel type, whether the current owned is an individual or a dealer and the type of transmission.



Year

2002

What is the Showroom Price?(In lakhs)

2

How Many Kilometers Drived?

40000

How much owners previously had the car(0 or 1 or 3)?

1

What Is the Fuel type?

Diesel

Are you A Dealer or Individual

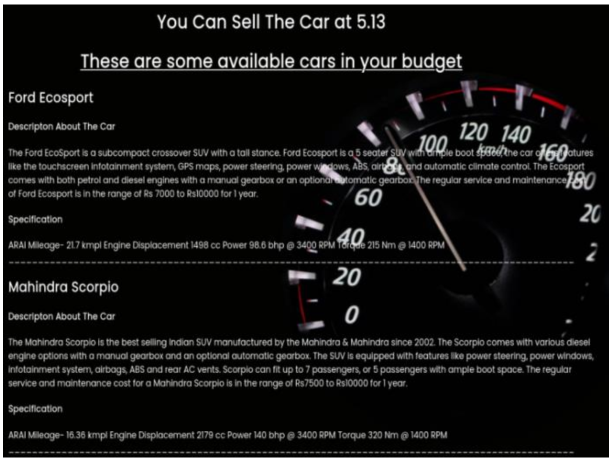
Dealer

Transmission type

Manual Car

Fig 4.1 User Interface to input data

Based on the given input, it gives the prediction which includes the selling price of the car and also some of the available cars in that price. The output of the predictive analysis is shown in Fig. 4.2.



You Can Sell The Car at 5.13

These are some available cars in your budget

Ford Ecosport

Descripton About The Car

The Ford EcoSport is a subcompact crossover SUV with a tall stance. Ford Ecosport is a 5 seater SUV with ample boot space, the car offers features like the touchscreen infotainment system, GPS maps, power steering, power windows, ABS, airbags, and automatic climate control. The Ecosport comes with both petrol and diesel engines with a manual gearbox or an optional automatic gearbox. The regular service and maintenance cost of Ford Ecosport is in the range of Rs 7000 to Rs10000 for 1 year.

Specification

ARAI Mileage- 21.7 kmpl Engine Displacement 1498 cc Power 98.6 bhp @ 3400 RPM Torque 215 Nm @ 1400 RPM

Mahindra Scorpio

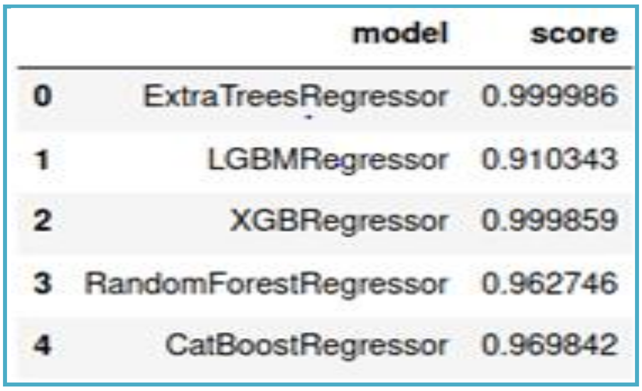
Descripton About The Car

The Mahindra Scorpio is the best selling Indian SUV manufactured by the Mahindra & Mahindra since 2002. The Scorpio comes with various diesel engine options with a manual gearbox and an optional automatic gearbox. The SUV is equipped with features like power steering, power windows, infotainment system, airbags, ABS and rear AC vents. Scorpio can fit up to 7 passengers, or 5 passengers with ample boot space. The regular service and maintenance cost for a Mahindra Scorpio is in the range of Rs7500 to Rs10000 for 1 year.

Specification

ARAI Mileage- 16.36 kmpl Engine Displacement 2179 cc Power 140 bhp @ 3400 RPM Torque 320 Nm @ 1400 RPM

Fig 4.2 Predictive Analysis data (OUTPUT)



	model	score
0	ExtraTreesRegressor	0.999986
1	LGBMRegressor	0.910343
2	XGBRegressor	0.999859
3	RandomForestRegressor	0.962746
4	CatBoostRegressor	0.969842

Fig. 4.3 Accuracy of various models

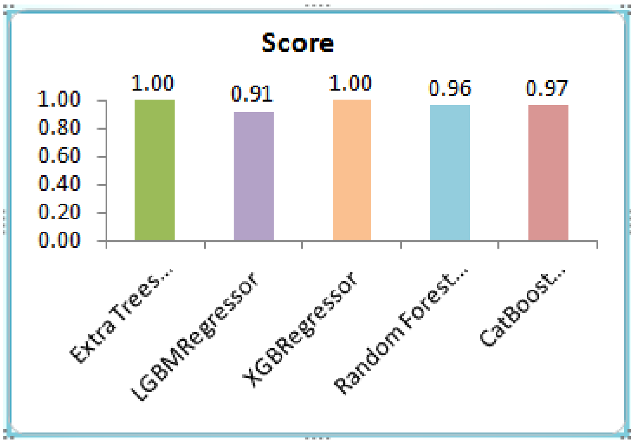
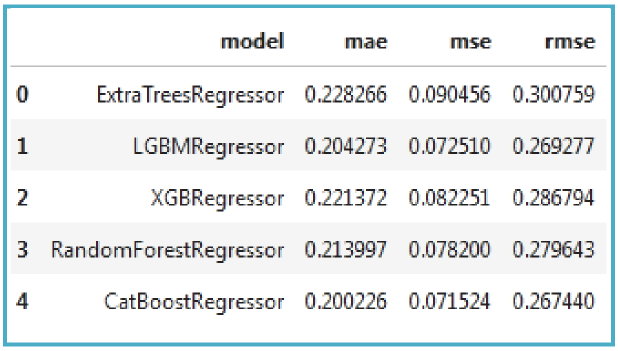


Fig. 4.4 Accuracy obtained from various models

The score or accuracy obtained from the process of applying various machine learning models to the dataset is shown in Fig. 4.3 and 4.4.



	model	mae	mse	rmse
0	ExtraTreesRegressor	0.228266	0.090456	0.300759
1	LGBMRegressor	0.204273	0.072510	0.269277
2	XGBRegressor	0.221372	0.082251	0.286794
3	RandomForestRegressor	0.213997	0.078200	0.279643
4	CatBoostRegressor	0.200226	0.071524	0.267440

Fig. 4.5 MAE, MSE and RMSE of various models

The MAE, MSE and RMSE (the most commonly used metrics to measure the accuracy for continuous variables) of various models are as shown in Fig. 4.5.

V. CONCLUSION

This project work can be concluded with the comparable results of both feature selection algorithms and classifier. This combination has achieved maximum accuracy and selected minimum but most appropriate features. It is important to note that in forward selection by adding irrelevant or redundant features to the data set decreases the efficiency of both classifiers. While in backward selection if we remove any important feature from the data set, its efficiency decreases. The main reason of low accuracy rate is low number of instances in the data set. Since CatBoost regressor has the lowest Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) it is suggested as the best model.

VI. REFERENCES

[1] Sameerchand Pudaruth, "Predicting the Price of Used Cars using Machine Learning Techniques", IJICT 2014.

[2] Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric, "Car Price Prediction using Machine Learning Techniques", TEM Journal-2019.

[3] Mariana Listiani, "Support Vector Regression Analysis for Price Prediction in a Car Leasing Application", Master Thesis, Hamburg University of Technology.

[4] Feng Wang; Xusong Zhang; Qiang Wang "Prediction of Used Car Price Based on Supervised Learning Algorithm", International Conference on Networking, Communications and Information Technology, 2021.

[5] Prediction of prices for used car by using regression models, 5th International Conference on Business and Industrial Research, 2018.