PERFORMANCE EVALUATION OF SOME MACHINE LEARNING ALGORITHMS

A. A. Ibrahim[1], O. A. Ayilara-Adewale[2], A. A. Alabi[3], F. R. Olokun-Olukotun[4], O. N. Ajadi[5] & O. T. Ogundele[6],

[1, 5,6]Department of Mathematical Sciences, Oduduwa University, Ipetumodu, Ile-Ife, Nigeria
[2] Department of Information Technology, Osun State University, Osogbo
[3] Department of Information Systems, Osun State University, Osogbo
[4] Department of Software Engineering, Osun State University, Osogbo
Corresponding author: adebisiibrahim97@gmail.com

Received: February 14, 2025, Accepted: April 28, 2025

Abstract: Stock market price prediction offers investors an insight into the dynamics of price fluctuations and how to make informed decisions. The coming of technology has even made price prediction more robust and accurate. In this paper, two models of machine learning, the Random Forest and Extra Tree models, were discussed. Data obtained from Kaggle.com were preprocessed. The features used in car price prediction included type, brand, number of years used, number of doors, body type, fuel type, and whether it has been registered. These are the most common features used in buying cars in developing countries like Nigeria. The models were developed using machine learning algorithms and implemented using Python and Scikit-learn machine learning libraries. Their performances were evaluated using the mean square error and R-squared. Though Random Forest and Extra Tree models showed similar results 91% and 92% mean squared errors respectively, the extra tree model had a lower prediction error of 7.5 as against 8.7 prediction error of the random forest model.

Keywords: Machine Learning, Random Forest, Extra Tree, Used Car, Price Prediction.

Introduction

Stock market operators endeavour to look into the future prices of commodities ranging from housing to automobiles, equipment and machineries, electronic gadgets, etc. to enable them to make informed decisions on their investments with the hope of optimizing their profits (Bremen, 2000). Traditional methods of price estimation of commodities often rely on manual analysis and subjective judgment, which in most cases takes more time and are likely to be erroneous. Lately, machine learning approaches have gained popularity in the field of car pricing prediction, offering the potential for more accurate and efficient price predictions for commodities (Mamipour et al, 2015, Kumar *et al*, 2022, Sadia et al, 2022)). Accurately predicting prices of commodities has become necessary since it gives buyers insight into the available products and their respective prices as well as financial institutions that are likely to finance the purchase give their customers expert advice and also enabling sellers take proper decisions. Apart from its primary purpose, these commodities also serve as investments for their owners. Thus, a proper price prediction is of paramount importance.

With the advent of technology, machine learning has introduced another dimension to price prediction (Geurts et al, 2006, Lee, 2002) thereby making commodity appraisal more convenient and reliable (Ben, 2015, Adhikary et al, 2022). In this paper, we compare the performance of two stock market price prediction models; the Random Forest and the Extra tree algorithms with particular reference to used automobiles. This is as a result of increase in transport business in Nigeria of today.

The used car market started at the turn of the 20th century. The need for affordable transportation pioneered second-hand car sales (Smith, 2005). After World War II, the used car market experienced significant growth. Second hand vehicles such as military jeeps and trucks flooded the automobile market as soldiers returned from war (Brown, 1999). Dealership in used car expanded, offering a wide range of models. Auto auctions became popular and provided a platform for used car transactions between buyers and sellers. (Johnson, 2010). In the mid-20th century, quality concerns arose due to unscrupulous practices and consumer protection laws were enacted to regulate the industry and ensure fair transactions (Garcia, 1987). The coming of digital revolution in the late 20th century ushered in online platforms like Auto-Trader transforming the used car market and allowing buyers to search and compare listings easily (Lee, 2002). Environmental awareness has also impacted the used car market. Intending car owners now put into consideration the fuel efficiency, emissions and other factors when buying second hand vehicles (Green, 2015).

MATHERIALS AND METHODS

The dataset for this work was the used cars dataset from Kaggle.com, which was preprocessed to find correlations between the features and the target (car price) for training the Random Forest and Extra Trees regression models. Algorithms for the models were developed and the models were designed and implemented using the Python programming language.

Random Forest and Extra Tree are two ensemble techniques that are similar in many ways but their differences lie in the fact that Extra Trees sampling is without replacement and its nodes are randomly split at a higher degree of tree construction than Random Forest (Geurts et al, 2006).

Random Forest as an ensemble learning technique simultaneously uses multiple decision trees for accurate predictions. It equally introduces randomness and diversity by training each tree on a random subset of data and features. handles missing data, is robust to outliers, and can be trained efficiently on large datasets and provides feature importance. Thus, it is versatile and suitable for complex data with high-dimensional features (Thamarai et al, 2020).

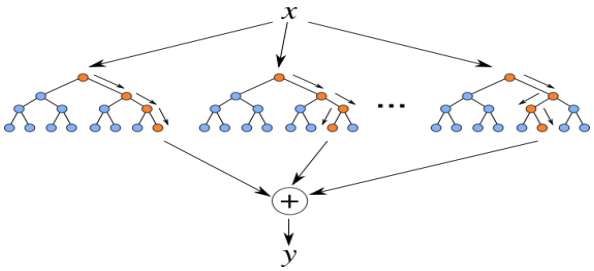


Figure 1. Random Forest

The Extra Trees (Extremely Randomized Trees) regression algorithm from another point of view is ensemble learning techniques that builds multiple decision trees and aggregates their results for improved predictive accuracy and robustness. The algorithm's key characteristics and design are detailed below:

1. Ensemble Method: Extra Trees is an ensemble method that constructs more decision trees during training. The different training data subsets the sources of building each tree. The mean of all the predictions of all trees is then used to obtain the final prediction.
2. Random Splits: Unlike traditional decision tree algorithms that find the best split at each node, Extra Trees selects randomly a subset of features and chooses the best split

among them. This increases the diversity of the trees and helps in reducing overfitting.

3. Bootstrap Aggregating: Extra Trees uses bootstrap aggregating (bagging) where each tree is trained on a random sample of the data with replacement. This technique reduces variance and helps in building a more generalizable model.
4. Feature Selection: For each node in a tree, a random subset of features is chosen, and the best split is determined based on these features. This randomness helps in capturing different aspects of the data and improves model performance.
5. Handling High-Dimensional Data: Extra Trees can handle high-dimensional data efficiently by considering only a subset of features for splitting at each node, thus reducing computational complexity.
6. Parallelization: The algorithm supports parallel processing, allowing multiple trees to be built simultaneously, which speeds up the training process (Fadzilah et al, 2021, Zhang et al, 2017).

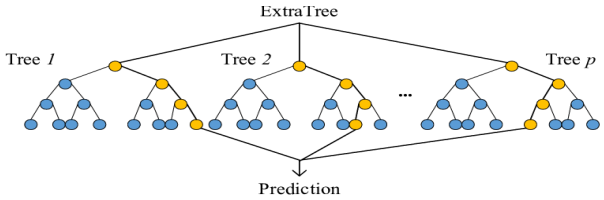
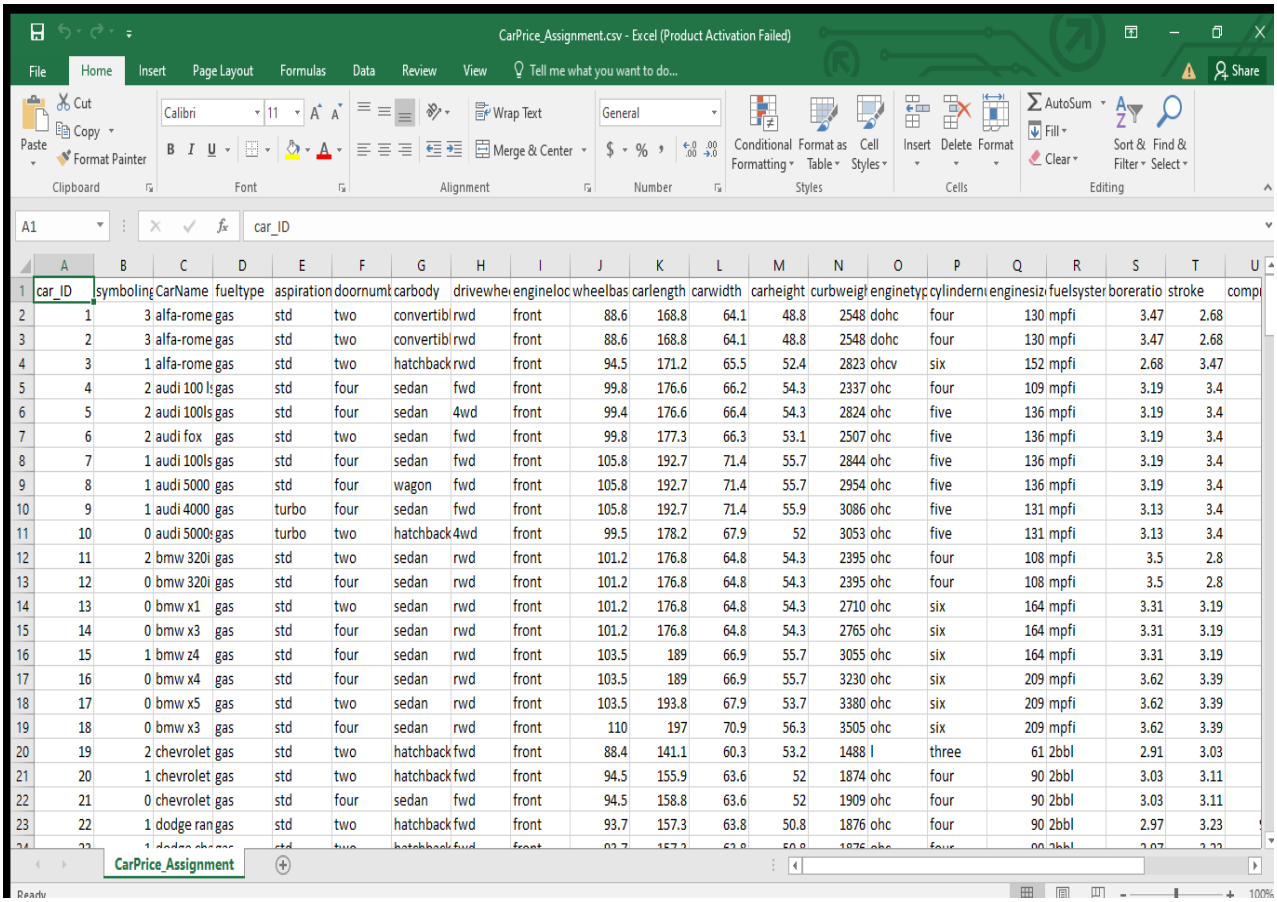


Figure 2: Extra Tree



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	car_ID	symboling	CarName	fueltype	aspiration	doornumb	carbody	drivewhe	enginelo	wheelbas	carlength	carwidth	carheight	curbweigh	enginety	cylinderni	enginesiz	fuelsyster	boreratio	stroke	compi
2	1	3	alfa-rome	gas	std	two	convertibl	rwd	front	88.6	168.8	64.1	48.8	2548	dohc	four	130	mpfi	3.47	2.68	
3	2	3	alfa-rome	gas	std	two	convertibl	rwd	front	88.6	168.8	64.1	48.8	2548	dohc	four	130	mpfi	3.47	2.68	
4	3	1	alfa-rome	gas	std	two	hatchback	rwd	front	94.5	171.2	65.5	52.4	2823	ohcv	six	152	mpfi	2.68	3.47	
5	4	2	audi 100 l	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.3	2337	ohc	four	109	mpfi	3.19	3.4	
6	5	2	audi 100ls	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.3	2824	ohc	five	136	mpfi	3.19	3.4	
7	6	2	audi fox	gas	std	two	sedan	fwd	front	99.8	177.3	66.3	53.1	2507	ohc	five	136	mpfi	3.19	3.4	
8	7	1	audi 100ls	gas	std	four	sedan	fwd	front	105.8	192.7	71.4	55.7	2844	ohc	five	136	mpfi	3.19	3.4	
9	8	1	audi 5000	gas	std	four	wagon	fwd	front	105.8	192.7	71.4	55.7	2954	ohc	five	136	mpfi	3.19	3.4	
10	9	1	audi 4000	gas	turbo	four	sedan	fwd	front	105.8	192.7	71.4	55.9	3086	ohc	five	131	mpfi	3.13	3.4	
11	10	0	audi 5000	gas	turbo	two	hatchback	4wd	front	99.5	178.2	67.9	52	3053	ohc	five	131	mpfi	3.13	3.4	
12	11	2	bmw 320i	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2395	ohc	four	108	mpfi	3.5	2.8	
13	12	0	bmw 320i	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2395	ohc	four	108	mpfi	3.5	2.8	
14	13	0	bmw x1	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2710	ohc	six	164	mpfi	3.31	3.19	
15	14	0	bmw x3	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2765	ohc	six	164	mpfi	3.31	3.19	
16	15	1	bmw z4	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.7	3055	ohc	six	164	mpfi	3.31	3.19	
17	16	0	bmw x4	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.7	3230	ohc	six	209	mpfi	3.62	3.39	
18	17	0	bmw x5	gas	std	two	sedan	rwd	front	103.5	193.8	67.9	53.7	3380	ohc	six	209	mpfi	3.62	3.39	
19	18	0	bmw x3	gas	std	four	sedan	rwd	front	110	197	70.9	56.3	3505	ohc	six	209	mpfi	3.62	3.39	
20	19	2	chevrolet	gas	std	two	hatchback	fwd	front	88.4	141.1	60.3	53.2	1488	l	three	61	2bbl	2.91	3.03	
21	20	1	chevrolet	gas	std	two	hatchback	fwd	front	94.5	155.9	63.6	52	1874	ohc	four	90	2bbl	3.03	3.11	
22	21	0	chevrolet	gas	std	four	sedan	fwd	front	94.5	158.8	63.6	52	1909	ohc	four	90	2bbl	3.03	3.11	
23	22	1	dodge ran	gas	std	two	hatchback	fwd	front	93.7	157.3	63.8	50.8	1876	ohc	four	90	2bbl	2.97	3.23	

Figure 3. Used Cars Dataset

Dataset which contains the variables employed in prediction of used cars price as well as libraries like pandas, NumPy and matplotlib, that aids data pre-processing and training of the model were imported and the systems were implemented using python programming la

Relevant data of used cars was collected from Kaggle.com for the training and testing of the model.

The required training dataset for this project was obtained from Kaggle.com. The attributes taken into account for car price prediction included type, brand, number of years used, number of doors, body type, fuel type, and whether it has been registered which are the most widely used parameters for buying a used car in developing nations.

All these features in the used cars dataset capture different aspects of the vehicles and provide insight into the factors that influence car prices. Understanding these features helps in developing a regression model and making informed decisions regarding car pricing and investments. The models were developed using Random Forest and Extra Tree machine learning algorithms and flowcharts. Thereafter, the algorithms were implemented using Python and Scikit-learn machine learning libraries and evaluated using mean squared error and R^2 score.

```
df = pd.read_csv('Data.csv')
print(df.shape)
df.head()
```

Python

```
(506, 14)
```

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2

Figure 4. Reading Dataset

And the summary statistics is depicted in figure 4

```
features = df.drop('MEDV', axis = 1)
features.head()
```

[3] Python

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33

Figure 5. Summary Statistics

The systems were evaluated using the mean square error (MSE) and R-squared (R^2). MSE measures the mean of the squared difference between actual and predicted values, with larger errors more penalized.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(i - \hat{Y}_i)^2$$

- n: total observations.
- i: individual observation.
- $(i - \hat{Y}_i)^2$: squared difference between the predicted value i and the actual valueY_i.

The R^2 score is the measure of how well the regression line approximates the actual data, i.e. goodness of fit of a model where

$$R^2 = 1 - \frac{SS_R}{SS_M}$$

- SS_R is the sum of squared error by regression line
- SS_M represents the sum of squared error by mean line.

RESULTS AND DISCUSSION

The minimum requirements for the system are:

a) Operating System
   i. Windows 7 or above
   ii. Mac OS 10.5.8 or later
   iii. Linux

b) Minimum processor and RAM
   i. 2.3GHz Pentium D processor
   ii. 4GB RAM
   iii. Graphics Card (optional)

The performance evaluation metrics for the random forest and extra tree prediction models on the test set were performed using MSE and R^2 score.

The models were evaluated with random forest model having $MSE = 8.7$ and $R^2 = 0.91$ and the extra tree prediction model having $MSE = 7.5$ and $R^2 = 0.92$

The differences between the predicted and actual prices of used cars were calculated. The results are presented in Table 1 below.

Table 1. Actual vs predicted Car prices

	Random Forest Method			Extra Tree Method		
S/N	Actual Value	Predicted Value	Difference	Actual Value	Predicted Value	Difference
1	30760	35854.085	-5094.09	30760	30482.7	277.29
2	17859.2	18938.93	-1079.76	17859.2	20197.9	-2338.7
3	9549	9021.8	527.2	9549	8938.01	610.09
4	11850	12894.8	-1044.8	11850	12895.9	-1045.9
5	28248	26908.96	1339.04	28248	28758.3	-510.29
6	7799	6506.24	1292.76	7799	6454.69	1344.31
7	7788	7762.08	25.92	7788	7716.9	71.1
8	9258	7990.34	1267.66	9258	7822.65	1435.35
9	10198	9875.34	322.66	10198	9949.57	248.43
10	7775	8301.855	-526.855	7775	8030.19	-255.19
11	13295	14157.18	-862.18	13295	15157.9	1862.9
12	8238	7858.89	379.11	8238	7612.33	625.67
13	18280	13555.17	4724.83	18280	16114.3	2165.73
14	9988	10722.95	-734.95	9988	11135.7	-1147.7
15	409060	39655.105	1304.895	409060	43746.6	-2786.6

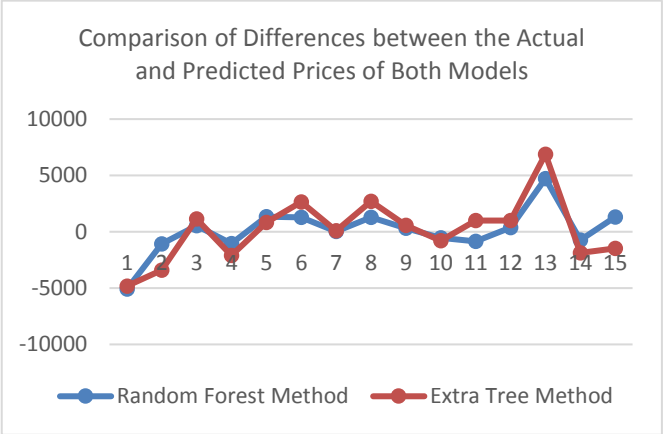


Figure 6. Graph of the Errors

Conclusion

This study compared two machine learning algorithms; the random forest and extra tree algorithms employing the same data set from Kaggle.com. The robustness of the two models is evident in the results obtained. While both the random forest and extra tree regression models predicted more accurately with almost the same mean square error (MSE) of 91% and 92% respectively, their R^2 values showed that the extra tree model has a slightly lower prediction error of 7.5 as against the 8.7 prediction error of random forest. The Extra Tree algorithm has a higher degree of randomization in constructing the trees which in turn makes for its

faster training. This we can say accounted for a better performance. This is equally evident from the graphical comparison of their respective errors. This makes the extra tree model more accurate than the random forest model.

Reference

Smith, J. 'History of Used Cars'. *Journal of Automotive History*, 20(3) 2005, 45-58.

Brown, R. 'Post-World War II Expansion of the Used Car Market'. *Economic Review*, 15(2), 1999, 112-125.

Johnson, M. 'Evolution of Auto Auctions in the Used Car Market.' *Auction Insights*, 5(4), 2010, 225-238.

Garcia, L. 'Consumer Protection Laws and the Used Car Industry.' *Legal Studies Journal*, 1987, 12(1), 30-42.

Lee, S. 'Digital Transformation of the Used Car Market'. *Online Marketplaces*, 8(3), 135-148.

Green, P. 'Environmental Influence on the Used Car Market'. *Technologies Review*, 25(2), 78-91.

Breiman, L. 'Randomizing Outputs to Increase Prediction Accuracy' *Machine Learning*, Vol. 40, No. 3, 2000

Mamipour, S. and Vaezi, F. 'Non-Linear Relationships Among Oil Price, Gold Price and Stock Market Returns in Iran: A Multivariate Regime-Switching Approach'.*Iranian Journal of Economic Studies*. Vol.4, No. 1, 2015.

Kumar, P. D., Sivadeep, K., Pavan, J. and Navaneeth, J. 'Stock market prediction using machine learning and Python'. *Journal of Financial Markets*, Vol 10, No. 1, 2022

Sadia, K. H., Sharma, A. Paul, A., Padhi, S. and Sanyal, S. 'Stock market price prediction of used cars using machine learning algorithms' *International Journal of Innovative Technology and Exploring Engineering*. Vol. 8, No. 11, 2022.

Geurts, P., Ernst, D. and Wehenkel, L. 'Extremely Randomized Trees' *Springer Science + Business Media, LLC* Vol. 6, No. 3, 2006

Ben J, 'Hedonic Regression Models for Used Car Appraisal'. *Appraisal Journal*, 18(2), 95-108.

Adhikary, D. R., Sahu, R. and Panda, S. P. 'Machine Learning Models for Used Car Price Prediction'. *Journal of Machine Learning Research*, 40(4), 2022, 300-315.

Thamarai, M. and Malarvizhi, S. P. 'Data Pre-processing and Decision Tree Models for Accurate Predictions'. *Journal of Computer Science and Information Technology*. Vol. 10 No. 2, 2020.

Fadzilah, S. and Abu, N. A. 'Used Car Price Estimation: Moving from Linear Regression Towards A New S-Curve Model'. *International Journal of Business and Society*. Vol. 22, No. 3, 2021

Zhang, L., Aggarwal, C. and Qi, G. 'Stock Price Prediction via Discovering Multi-Frequency Trading Patterns, (ACM SIGKDD). presented at the *23rd International Conference on Knowledge Discovery and Data Mining, Halifax, Nova Scotia, Canada*, 2017.

Ho, T. K. 'Random Decision Forests'. *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC*, 14–16 August 1995. pp. 278–282. Archived from the original (PDF) on 17 April 2016.