Forecasting Vehicle Prices using Machine Learning Techniques based on Federated Learning Strategy

Mohammed A. Mahfouz
Assistant Lecturer at Thebes Academy, Information System Department, Cairo, Egypt

Sara M. Mosaad
Lecturer at Faculty of Commerce, Business Information System Department, Helwan University, Cairo, Egypt

Mohamed A. Belal
Prof. of Computer Science Department, Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt

ABSTRACT

With the growth of the Internet and big data, more consumer behavior data is being used in various forecasting issues, enhancing prediction accuracy. Due to market and environmental changes, automobile sales will fluctuate as the primary mode of transportation. Accurate car sales forecasting can influence the economy and the transportation industry and enable dealers to dynamically alter their marketing strategies. A variety of factors can influence the decision to buy a vehicle, including the product's inherent characteristics, economics, policy, and other factors. Additionally, the sample data display traits from various sources, tremendous complexity, and high volatility.

To estimate the monthly sales of autos, this study employs a variety of machine learning and statistical models, each of which has global optimization, a simple structure, and high generalization capabilities. Additionally, Federated Learning is perfecting the parameters to boost data security and prediction accuracy. The elements that influence auto sales are first examined and determined using statistical and machine approaches. Third, we use the dataset for the experimental analysis using the Dickey-Fuller test. The findings show that the Fed-Linear Regression model shows the best mean absolute percentage error (MAPE) and root-mean-square error (RMSE) performance. Finally, managerial implications are put forward for reference.

General Terms

XGBoost Regression, Random Forest Regression (RFR), Linear Regression (LR), Machine learning.

Keywords

Fed-KNN, Fed-SVM, Fed-Random Forest, Fed-Neural Network, Fed-Linear Regression, Fed-Gradient Boosting, Fed-AdaBoost.

1. INTRODUCTION

Commodities have long been a part of lifestyles because they are a continuous paradigm of transactions. Earlier, those exchanges took the shape of a bartering device an intermediary, which was later transformed into an economic device. And taking all these changes into account, the way items were re-marketed underwent a similar transformation. There are methods where the object's resale is completed [1].

One is offline, while the other is live. In offline transactions, an intermediary is highly likely to be dishonest and make excessively lucrative deals. The second choice is to advertise online, where there may be a specific platform where a person can discover what they might pay. Kilometers driven - When placing a car up for sale, the number of kilometers driven with its help is especially important. The car's kilometers get older the more it has driven. Fiscal energy is the amount of power a car produces. Higher output results in a higher expense for an automobile. [1]

The year the car was registered with the Road Transport Authority is known as the "year of registration." The value of a car will increase as it becomes more modern. The value of the car will decrease with each passing year. Fuel Type: The data collection that will be used in this paper has two different fuel types: gasoline and diesel. To forecast the price of used cars, a machine-learning or self-learning capable system is needed. The primary goal of this paper is to develop a supervised machine-learning model for forecasting the value of a car based on multiple attributes, which will be a real-time job [2].

Predicting car prices is very intriguing and well-known. The market for used cars has shown a sharp increase in worth, which has increased its market share. In India, there are close to 3.4 million used cars sold annually.

This gives car price prediction even more significance. Because rates typically depend on a variety of factors, correct car price prediction requires expert knowledge. Typically, the most important ones are the brand, model, age, horsepower, mileage, and economic variables affecting demand and supply[2].

Due to the frequent increases in the cost of fuels such as diesel, gasoline, and compressed natural gas, the gasoline used in a car has an impact on the price of an automobile. Other technical, agronomical, and figurative features such as exterior color, door variety, type of transmission, size, passenger safety features such as airbags and air conditioning, interior features such as upholstery, navigation and infotainment systems, cruise controls, and so forth may also have an impact on car prices [2].

2. RELATED WORK

As a result, increasing numbers of studies use information about consumer behavior to make predictions in various areas explanation studies on product sales exist. A better star rating increases the likelihood that a user will select a product, according to research by Kostyra et al. [3] on the correlation between online reviews and sales of e-book readers. According to Kulkarni et al. [4], the network search conducted before the release of a film adheres to a set of rules, and the model used in conjunction with the network search can greatly increase the predictability of the opening week box office. By examining the connection between computer sales, valence, and word-of-mouth emotional scores using log-linear regression.

Li et al. [5] discovered that valence moderated the impact of emotion on sales. After an auto recall, Wang et al. [6] addressed the combined effects of rivalry and social media activity on offline auto sales. On the other hand, there are predictive studies on product sales. A hybrid ISCA-BPNN model based on a Google trend was suggested by Hu et al. [7] to forecast the opening price direction and the future financial return. The technical indexes of stock prices, the text sentiment in news stories, and user data from social media platforms were combined by Li et al. [8] and Liu et al. [9] to forecast market movements. Google trends and transmission data linked to COVID-19 were used by Prasanth et al. [10] to forecast upcoming infections. By incorporating the Baidu index and Twitter's social media statistics.

Ai et al. [11] prediction of tourist flow was significantly improved. To accurately forecast automobile sales, Fan et al. [12] used the Bass diffusion model to replace the model coefficient in innovation diffusion with text emotion value. As can be seen, buyer behavior data are frequently used and perform well in a variety of product sales forecasting issues, including those involving the box office for movies [13], travel demand [14], the financial market [15], etc.

A product with excellent value and high participation is the automobile. Consumers must have a thorough grasp of the ideal model before they can make a purchase. Physical stores, product manuals, networks, expert forums, and others are the sources of this data. However, as an automaker, it only advertises its benefits to customers, which can create an informational imbalance and make it challenging for customers to get first-hand experience.

Online word-of-mouth information is currently a useful resource. Through a study of the current automobile forums, it discovered that each comment includes a star rating that customers have given for various aspects of the car. Customers' star ratings can accurately depict their opinions and unbiased assessments of the vehicle, which helps forecast sales of automobiles.

The major techniques for predicting sales, according to the literature currently available, include the time series method, linear regression, and machine learning method, gray forecasting method, and others. Grey prediction and autoregressive moving average are popular forecasting techniques for predicting auto sales. Shakti et al. [16] used the autoregressive integrated moving average (ARIMA) model to predict the trend of car sales.

Sa-ngasoongsong et al. [17] used the statistical unit root, weak exogenous, granger causality, and hull cointegration test to identify the dynamic coupling relationship between sales and economic indicators, and they established the vector error correction model for multi-model automobile sales. They did this while considering the influence of non-linear and non-stationery lines on automobile sales. To forecast car sales in Greece.

Konstantakis et al. [18] developed a vector autoregressive model. Using a generalized impulse response function, they calculated the long-term effects of various variables on car sales. Additionally, Ding and Li [19] generated a dynamic weighted sequence to highlight the new data without information omission and suggested an adaptive optimized gray model to predict the sales of electric cars. He et al. [20] suggested an optimized gray buffer operator by introducing the cumulative translation transformation and determining its optimal parameters by genetic algorithm to forecast the sales of new energy vehicles in China. To forecast the sales of new energy vehicles, Ding et al. [21] suggested an adaptive data pre-processing and optimized nonlinear gray Bernoulli model Mathematics 2022, 10, 2234 3 of 22 and demonstrated the high forecasting accuracy of this model.

The analysis of the current automobile industry reveals that the sales data for cars shows a clear non-linear trend of change. Social factors like word-of-mouth recommendations, Internet search volume, customer confidence levels, and other self-attributes like car color, material, and price are the primary determinants of auto sales. Additionally, the external environment, which includes the economy, raw materials, and policies, can influence car sales degree. A variety of variables influence the sales of automobiles. However, because the traditional prediction model struggles to manage multi-source data and correctly spot nonlinear relationships, machine learning-based prediction research is gradually advancing.

As machine learning is studied, its benefits in solving prediction problems start to become clear, and more academics are starting to use machine learning techniques for prediction. Artificial neural networks (ANN), random forests (RF), support vector machines (SVM), and support vector regression are often used techniques (SVR).

To forecast stock prices, Chen et al. [22] suggested a hybrid model that combines extreme gradient boosting (XGBoost) and enhanced the firefly algorithm. Vijh et al. [23] employed ANN and RF technology to forecast the closing prices of various industries' firms the following day. To improve the performance of the backpropagation neural network. Peng and Xiang [24] suggested a traffic flow prediction model using a genetic algorithm (GA) (BPNN). To predict network traffic, Yang et al. [25] proposed a simulated annealing optimization ARIMA-BPNN model, and the experiment findings confirmed the efficacy of the approach. As can be seen, the machine learning approach has been extensively used to solve prediction issues in a variety of fields, and it has the potential to enhance model performance. However, there has been less study on this topic because data on auto sales is primarily time series data. To predict auto sales, this paper makes use of machine learning techniques. It does so to explore the benefits of such techniques in the prediction of auto sales, as well as to offer more method options and pointers to expert opinions for interested researchers.

The proposed method is based on establishing the federated learning infrastructure, which includes a central server and multiple client devices. Each client device represents a data holder, such as a dealership or an individual seller, that contributes data for training the model without sharing it directly. Distribute the initial model to the client devices. Each client trains the model using its own local data without sharing it with others. Employ federated learning techniques, such as Federated Averaging, to aggregate the locally trained models and update the global model on the central server.

Monitor the performance of the model over time and collect new data to re-train and update the federated learning system periodically. This ensures that your model stays up to date and continues to provide accurate car price forecasts. As well as considering the privacy and security aspects associated with federated learning, as the data remains decentralized and protected on the client devices.

Table 1: The related work summary

Paper	Model
Kostyra et al. [3] - Kulkarni et al. [4] - Li et al. [5] - Wang et al. [6] - Hu et al. [7] - Li et al. [8] - Liu et al. [9] - Prasanth et al. [10] - Ai et al [11] - Fan et al. [12] - [13], [14], [15]	Correlation - Linear Regression - A Hybrid ISCA- BPNN
Shakti et al. [16]	Autoregressive Integrated Moving Average (ARIMA)
Sa-Ngasoongsong et al. [17]	Statistical Unit Root
Konstantakis et al. [18]	Vector Autoregressive Model
Ding And Li [19]	Adaptive Optimized Grey Model
He et al. [20]	Optimized Grey Buffer
Ding et al. [21]	Optimized Nonlinear Grey Bernoulli Model
Chen et al. [22]	Extreme Gradient Boosting (Xgboost)
Vijh et al. [23]	ANN And RF
Peng And Xiang [24]	Genetic Algorithm (GA) (BPNN).
Yang et al. [25]	simulated annealing optimization ARIMA-BPNN

As a result, this paper suggests a machine learning-based federated learning prediction model that can manage data from multiple sources and capture nonlinear data changes. The format of this paper is as follows: The methods used in this research are described in Section 2. The car sales prediction model used in this study is presented in Section 3. The comparative analyzes and implementation in practice are done in Section 4. Practice-related consequences are provided in Section 5. Section 6 comes to a climax at the end.

3. PROPOSED MODEL

Federated learning is a machine learning approach that enables multiple devices or entities to collaboratively train a model without sharing their data directly with each other. This approach can be applied to time series data forecasting as well. In time series data forecasting, federated learning can be used to aggregate data from various sources and train a model that can make exact predictions for future time periods. The model is trained on the data that is available on each device or entity, without having to share the data across them. This ensures the privacy and security of the data while still achieving an elevated level of performance.

With federated learning, each device or entity sends its updated parameters to a central server, which aggregates them and returns an updated model to each participant. This process is repeated iteratively, allowing the model to learn from the collective data available on all devices or entities.

However, there are challenges associated with using federated learning for time series data forecasting. For example, the data on each device or entity may be heterogeneous, noisy, or incomplete, which can affect the accuracy of the model. Also, the communication overhead involved in exchanging model parameters and updating the central server can cause delays and bandwidth issues. Hence, proper synchronization of the parameters across all devices or entities is needed to ensure the convergence of the model.

Overall, the use of federated learning in time series data forecasting holds promise in enabling correct predictions while keeping data privacy and security. This paper focuses primarily on two features:

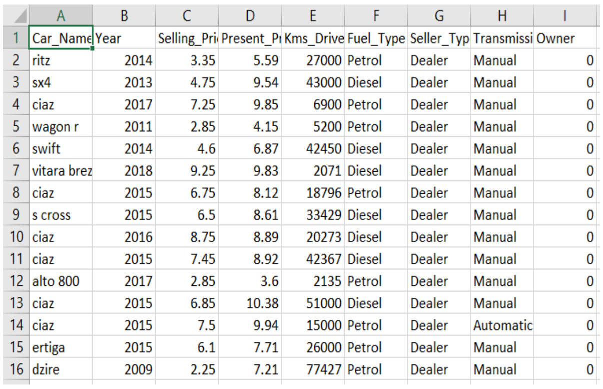
- Re-sale platform: A centralized platform for vehicle resale with price prediction.
- Feature selection: Prediction and search based on features.

The dataset is first gathered to begin the procedure. Data preprocessing, which includes data cleaning, data reduction, and data transformation, is the next stage. The price will then be predicted using a variety of machine learning techniques. Linear regression, Ridge regression, and Lasso regression are all used in the methods. The best model that accurately forecasts price is chosen. After choosing the best model, the user is shown the predicted price based on their inputs. On a website, users can supply data that will be used to forecast vehicle prices. To predict the price of a used car, this method compares the accuracy scores of Decision Tree, Logistic Regression, Random Forest, Gradient Boosting Algorithms, and Nave Bayes algorithms, among others. Figures 1 depicts the suggested system's block diagram and workflow.

Data was gathered from the Kaggle.com web portal, which offers the vehicle dataset from car Dekho for selling and buying cars. For each vehicle, the following characteristics were recorded:

Car Name, Year, Selling Price, Present, or the Current Price, Kilometers driven, Fuel Type: Petrol, Diesel or CNG (Compressed Natural Gas), Seller Type: Dealer or Individual, Transmission: Automatic or Manual, Owner (No. of earlier owners).

For this research, only the vehicles whose prices were listed in the dataset were listed and considered. Data with null entries were removed to keep the dataset's homogeneity because they could have had an impact on the prediction model. The column Car Name (Model) was also removed from the dataset due to the difficulty in finding safety data entries for the car name and model. The Car Name is also not extremely helpful in this research. Even though the data supported attributes were scant, their significance is still considered for this research.



	A	B	C	D	E	F	G	H	I
1	Car_Name	Year	Selling_Pri	Present_Pı	Kms_Drive	Fuel_Type	Seller_Typ	Transmissi	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
16	dzire	2009	2.25	7.21	77427	Petrol	Dealer	Manual	0

Fig 1: The sample of collected data.

3.1 XGBoost

With extreme gradient boosting, decision trees are produced consecutively. All independent factors are given weights, and the decision tree that forecasts outcomes uses these weights as input. The second decision tree is then given the variable weights that were incorrectly predicted. To create a model with greater accuracy, these separate predictors are combined.

The XGBoost (eXtreme Gradient Boosting) algorithm is a popular machine-learning algorithm that can be used for regression problems. The regression formula for XGBoost can be expressed as follows:

$$y = \sum m = 1mfm(x) \qquad (1)$$

Where y is the predicted value, M is the total number of trees in the ensemble (the total number of boosting iterations), fm(x) is the prediction made by the mth tree in the ensemble for a given input x.

Each tree in the ensemble is trained to minimize the residual error of the earlier trees. This iterative process helps to improve the accuracy of the model by reducing the bias and variance in the predictions. The XGBoost algorithm incorporates multiple techniques to improve the efficiency and accuracy of the model, including weighted quartile sketches, approximate greedy algorithms, and regularization parameters.

3.2 Decision Tree

The Extra Trees algorithm aggregates the results of multiple de-correlated decision trees collected in a forest to output its classification result. Here, each decision tree is built from the original training sample. Predictions are made by averaging the prediction of the decision trees with regression or using majority voting with classification.

3.3 Linear Regression

LR is used to predict the value of a variable based on the value of another feature. The feature you want to predict is called the dependent variable. The label that is used to predict the value of another feature is called the independent variable. The LR equation is of the form:

$$A = m + nB \qquad (2)$$

Where B is the independent variable, A is the dependent variable, A is the intercept y, and n is the slope of the line.

3.4 KNN Regression

KNN regression is a nonparametric technique that approximates the relationship between an independent variable and a continuous outcome by averaging observations in the same neighborhood. The analyst should specify the size of k. Alternatively, it can be chosen by cross-validation to choose the size that will minimize the mean squared error.

$$\text{Distance} = \sqrt{\sum_{i-1}^{k}(x_i - y_i)^2} \qquad (3)$$

3.5 Random Forest Regression

Random Forest Regression comes under a Supervised Machine Learning algorithm which uses an ensemble model for regression. Ensemble learning is a technique that clubs the outcomes of different machine learning models to create more predictions than one model. Random forest algorithm creates decision trees on data samples during the training and then gets the prediction from each of them. Finally, it selects the best solution by 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 correct prediction.

Random forest regression is a machine learning technique that uses a combination of multiple decision trees to predict the target variable. The formula for Random Forest Regression is:

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n \qquad (4)$$

Where, y is the dependent variable (target variable), x_1, x_2, ..., x_n is the independent variable (features) and b_0, b_1, b_2, ..., b_n = coefficients (parameters).

Random forest regression combines multiple decision trees, and each tree gives its prediction for the target variable. The final prediction is calculated by taking the average of all the predictions obtained from all the decision trees.

3.6 Bagging Regression

Bagging is short for Bootstrap Aggregating. It uses bootstrap resampling to train multiple models on random variations of the training set. At prediction time, each item's predictions are aggregated to give the final predictions. Bagged decision trees are efficient because each decision tree is suitable for a slightly different training data set. This allows each tree to have subtle differences and make slightly different skill predictions.

3.7 AdaBoost Regression

AdaBoost's model is a noticeably short single-level decision tree. Initially, it models a weak learner and gradually adds it to an ensemble. Each next model will try to change the predictions or outcomes of the earlier models in a series. It is achieved by weighting the train data to focus more on train examples in which old models made prediction errors.

$$y(x) = \frac{\sum_{t=1}^{r} \propto t * h_t(x)}{\sum_{t=1}^{r} \propto t} \qquad (6)$$

Where T is the number of iterations/weak models, Et is the weighted squared error of the t weak model, αt is the weight coefficient for the tth weak model, ht(x): the prediction of the tth weak model on example x and y(x) is the final prediction of the AdaBoost Regression model for example x.

3.8 ARIMA (Autoregressive Integrated Moving Average)

ARIMA is a model that combines the autoregressive, integrated, and moving average components to account for seasonality, trends, and noise in time series data. The autoregressive part models the relationship between an observation and number of lagged observations, while the moving average part models the relationship between observations and the residual errors. The integrated part takes the difference between consecutive observations to remove any trend from the data.

3.9 VAR (Vector Autoregression)

VAR is a model that considers the relationships between multiple time series variables. It models a set of related variables simultaneously, rather than modeling each variable separately as in ARIMA. VAR works by analyzing the correlation and causality among different time series variables, estimating their lagged effects on each other, and predicting their future values.

4. OPTIMIZING USING FEDERATED LEARNING

The approach is to develop a distributed application that calls for distributed data processing that is governed by resources and human factors. The central collection of data from each node might result in network traffic on a low-bandwidth wireless network, as well as power consumption. The communication burden and power consumption of diverse nodes in sensor networks are expected to be reduced by a distributed data mining design.

As a result, Data Mining Architecture must observe dispersed data, communication, calculations, and resources to make them. The goal of FL, as shown in Figure 4, is to undertake data mining activities based on the availability of resources and the type of operations. For data access, any site can be chosen, and activities can then be conducted.

4.1 Server-Side Model

In the server-side model, a set of machine learning models is being generated. Shop owners can initially log in to the system and enter their product details, which are kept in a database, by logging in. Additionally, the database has historical sales information. This data is introduced into the system and used to train the ML model. There are thousands of rows in the training data. Once the owner enters client wishes (the car color, price, velocity, etc.) the system will generate the machine learning model and transfer the client wishes and the code to the server side.

4.2 Client-side model

At the client-side model, the code sent from the server model will run on the set of datasets found on each client from the GPU's which are used in the model to be as different clients.

There is a little key benefit to using a GPU in parallel processing ends. First, modern GPUs are particularly well-suited to compute-intensive tasks that require enormous amounts of data to be processed. This is particularly true in areas such as deep learning and machine learning research, which require a huge amount of data, such as images or video, to be fed into a model for it to be trained. GPUs are also perfected for offloading data-intensive computations, as the hardware architecture of a GPU is more conducive to efficiently performing small calculations quickly.

The second benefit is the cost efficiency of GPUs. Given their greater power, you can typically get more bangs for your buck when compared to those of CPU solutions. GPUs offer significant speed improvements not only when working with high-performance tasks but also with simpler tasks, such as analyzing various large data sets in batches. This allows for faster iterations and better performance with less expense.

Finally, each client returns the results after applying the federated learning package (machine learning model) with the most suitable car that satisfies the client's wishes in this inventory.

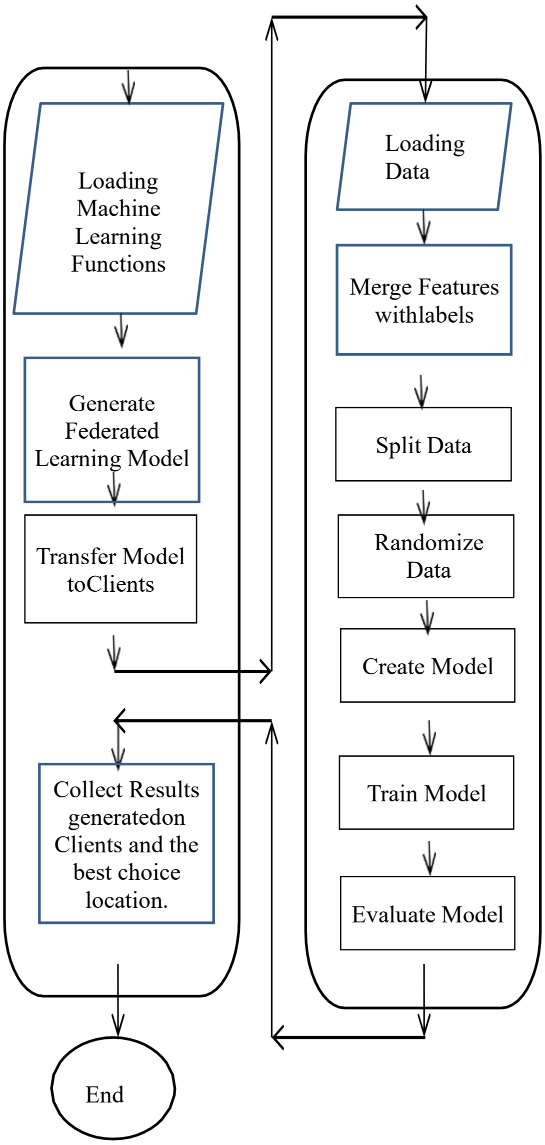

Loading Machine Learning Functions
Generate Federated Learning Model
Transfer Model toClients
Collect Results generatedon Clients and the best choice location.
End
Loading Data
Merge Features withlabels
Split Data
Randomize Data
Create Model
Train Model
Evaluate Model

Fig 2: The Proposed model flowchart

Table 2: The sample of the first dataset's statistical features.

Name	Mean	Mode	Median	Dispersion	Min.	Max.
Km driven	69819.51	120000	60000	0.81	1	2360457
seats	5.42	5	5	0.18	2	14
selling price	638271.81	300000	450000	1.26	29999	10000000
year	2013.8040	2017.0	2015.0	37.0 seconds	1983.0	2020.0

5. EXPERIMENTAL RESULTS

5.1 Datasets

For the experiment, the dataset used in the same area from different platforms objective is to test the relationship between customers and product sales to obtain future product sales forecasts [26], [27], [28], [29]. The validity of the proposed model is proved by using this case in this paper.

5.2 Evaluation Metrics.

Twenty machine learning models were evaluated in this study, including (Linear Regression (LR), Neural Network (NN), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), ADABoost, XGBoost (XB), Vector Autoregression (VAR), K-Nearest Neighbor (KNN), and Autoregressive Integrated Moving Average (ARIMA)). The tested mean squared errors (MSEs) and mean absolute errors (MAEs) obtained for each model on each dataset after applying the experiments on the preprocessed data using the prior models are shown in Table 4 and table 5.

MSE (Mean Squared Error) and RMSE (Root Mean Squared Error) are both measures of the difference between actual and predicted values. They are commonly used as evaluation metrics in regression analysis [30], [31], [32].

- Mean Squared Error (MSE):

It is the average of the squared differences between actual and predicted values. It is calculated using the following formula:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (yi - \hat{y}\mathrm{i})^2 \qquad (7)$$

Where n is the number of observations, ii is the actual value and ŷi is the predicted value.

- Root Mean Squared Error (RMSE):

It is the square root of the average of the squared differences between actual and predicted values. It is calculated using the following formula:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (yi - \hat{y}\mathrm{i})^2} \qquad (8)$$

Where n is the number of observations, y_i is the actual value and $\hat{y}_i$ is the predicted value. RMSE is a more popular metric than MSE, as it is in the same unit as the dependent variable, which makes it easier to interpret.

- MAE (Mean Absolute Error):

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |yi - \hat{y}\mathrm{i}\,| \qquad (9)$$

- R^2 (Coefficient of Determination):

$$R^2 = 1 - \frac{\sum_{i=0}^{n} (x_i - \hat{y})^2}{\sum_{i=0}^{n} (\hat{y} - \mathbf{Y_i})} \qquad (10)$$

R2 = 1 - (SSres/SStot) (11)

Where SSres = sum of squared residuals, SStot = total sum of squares.

- CVRMSE (Coefficient of Variation of Root Mean Squared Error):

$$\mathrm{CVRMSE} = (\mathrm{RMSE} / \text{mean of actual values}) * 100\% \qquad (12)$$

Figure 3, figure 4 and figure 5 are the prediction of automobile price for future 5 years using ARIMA, VAR and Neural Network, respectively. Figure 6 shows the ARIMA model for price prediction using federated learning on GPU (4 different processors).

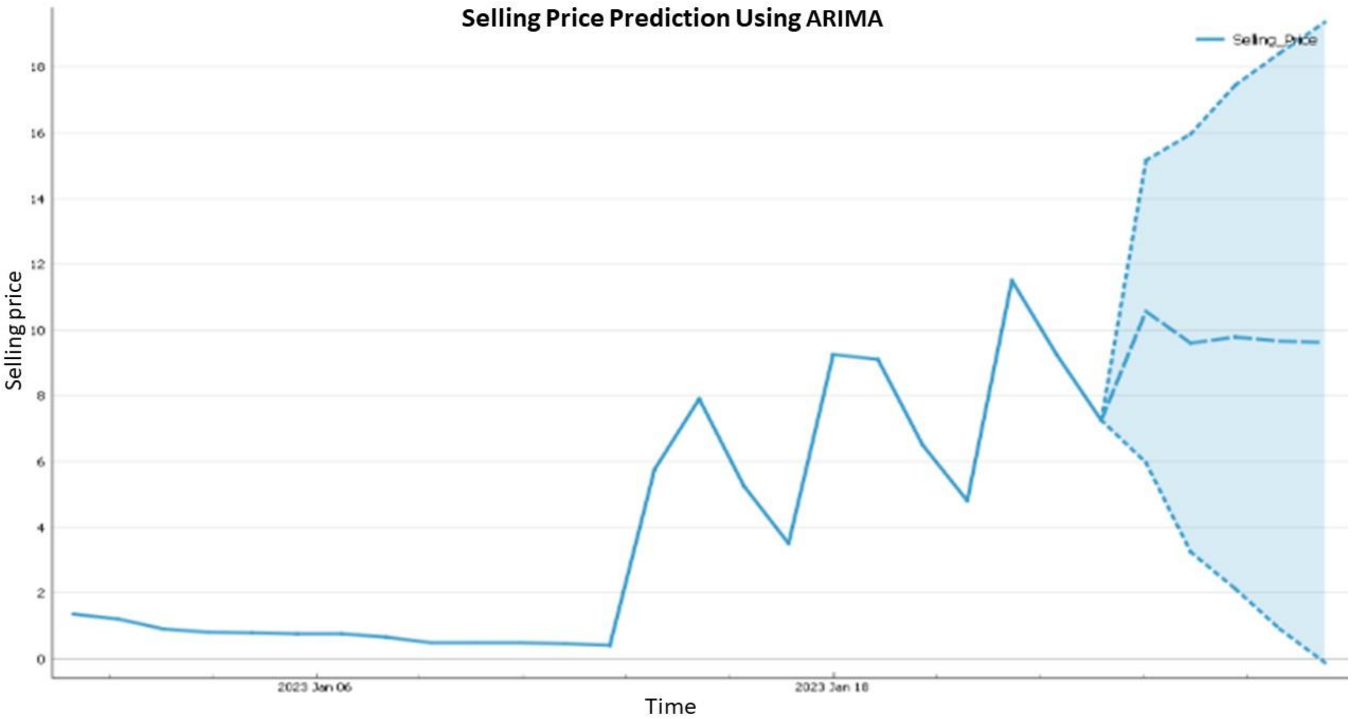


Fig 3: Selling price forecasting using ARIMA.

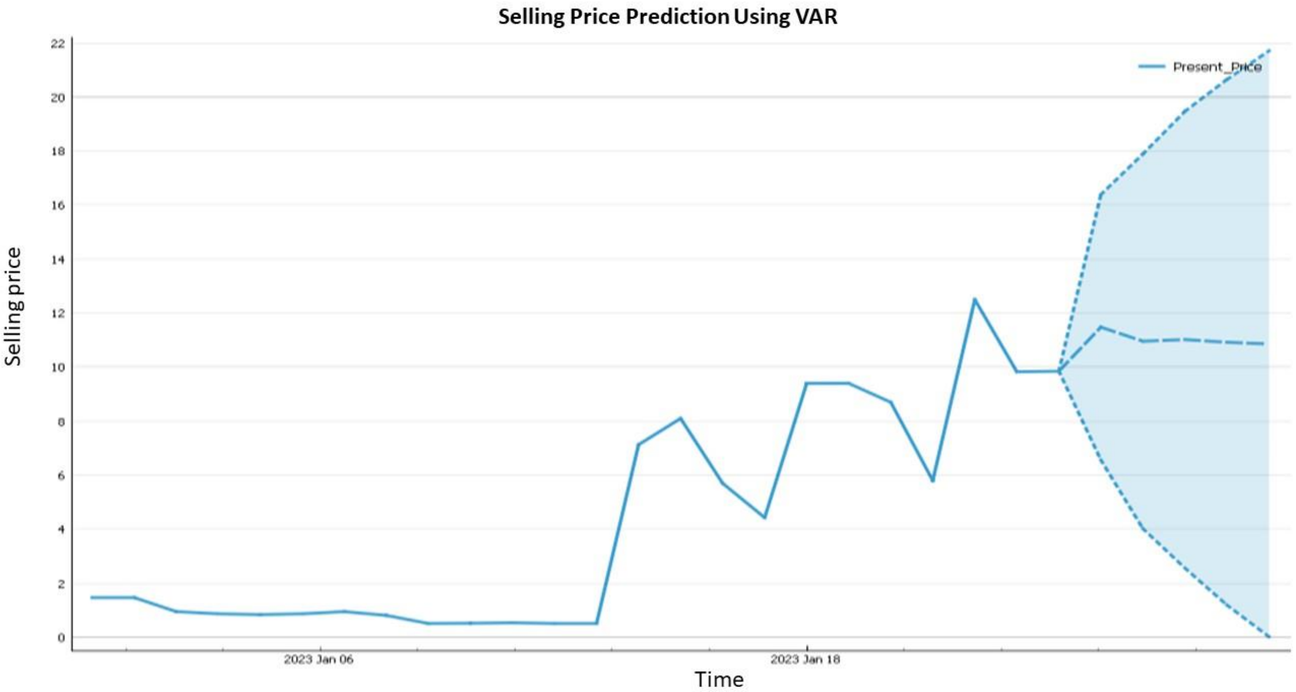


Fig 4: Selling price forecasting using VAR.

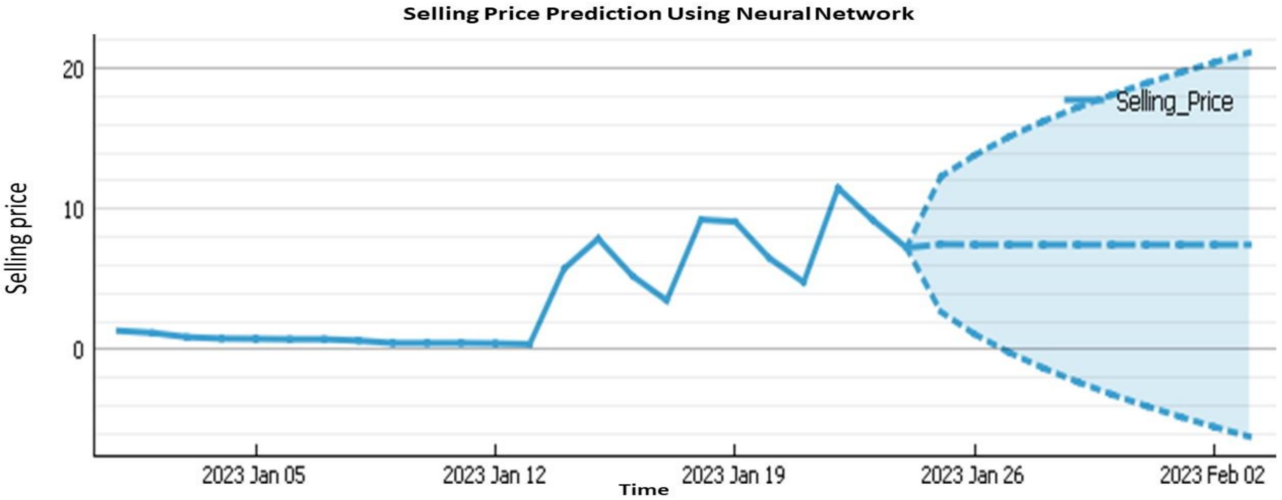


Fig 5: Selling price forecasting using the Neural Network.

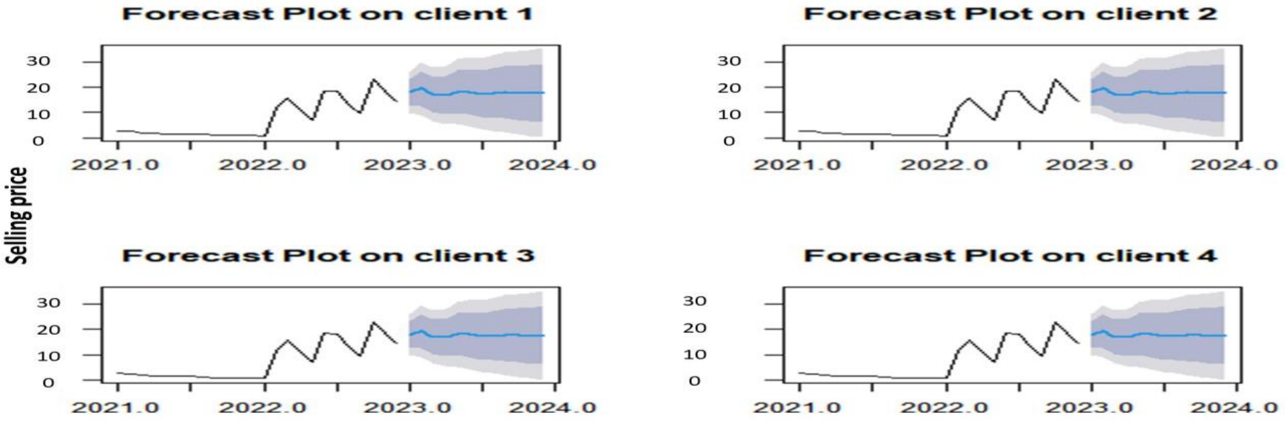


Fig 6: Selling price forecasting using federated learning with GPU.

Table 3: The forecasted selling price using ARIMA

selling price (forecast)		selling price (95%CI low)	selling price (95%CI high)
1	666660	-835960	2.16928e+06
2	647057	-925878	2.21999e+06
3	640991	-938515	2.2205e+06
4	639113	-941021	2.21925e+06
5	638532	-941662	2.21873e+06

Table 4 and table 5 present a comparison between the accuracy of machine learning models on automobile dataset in terms of MSE, RMSE, MAE, R2 and CVRMSE.

Table 4: Sample of machine learning models' accuracy in terms of MSE, RMSE, MAE, R2 and CVRMSE.

Model	Train time [s]	Test time [s]	MSE	RMSE	MAE	R2	CVRMSE
KNN	0.027	0.636	3.012	1.735	1.035	0.835	39.882
SVM	0.037	0.005	4.01	2.002	1.143	0.78	46.018
Random Forest	0.299	0.018	1.918	1.385	0.836	0.895	31.822
Neural Network	0.832	0.009	16.556	4.069	3.86	0.092	93.504
Linear Regression	0.026	0.004	0.634	0.797	0.527	0.965	18.304
Gradient Boosting	0.297	0.005	2.562	1.601	0.9	0.859	36.785
AdaBoost	0.686	0.05	2.529	1.59	0.891	0.861	36.544
ARIMA	0.027	0.636	3.012	2.34	0.899	0.452	46.018
VAR	0.037	0.005	4.01	1.437	1.346	0.623	39.882

Table 5: Sample of machine learning models accuracy in terms of MSE, RMSE, MAE, R2 and CVRMSE using Federated Learning.

Model	Train time [s]	Test time [s]	MSE	RMSE	MAE	R2	CVRMSE
Fed-KNN	0.012	0.004	0.835	0.914	0.566	0.936	24.505
Fed-SVM	0.019	0.005	2.376	1.542	0.811	0.818	41.327
Fed-Random Forest	0.253	0.013	1.5	1.225	0.743	0.885	32.839
Fed-Neural Network	0.052	0.009	0.292	0.541	0.352	0.978	14.495
Fed-Linear Regression	0.011	0.002	0.292	0.541	0.352	0.978	14.495
Fed-Gradient Boosting	0.259	0.004	1.728	1.314	0.722	0.868	35.241
Fed-AdaBoost	0.61	0.028	1.733	1.316	0.723	0.867	35.295
ARIMA	0.022	0.436	1.012	2.34	0.899	0.452	46.018
VAR	0.017	0.003	2.01	1.437	1.346	0.623	39.882

Table 6: Sample of Selling price (in hundred thousand) and different machine learning models predicted selling price values.

Linear Regression	Random Forest	AdaBoost	KNN	SVM	Neural Network	Gradient Boosting	Selling Price
0.414	0.569	0.48	0.568	0.728	4.57	0.465	0.4
0.391	0.527	0.4	0.558	0.728	4.235	0.4	0.45
0.391	0.527	0.4	0.558	0.728	4.235	0.4	0.48
0.423	0.569	0.48	0.568	0.729	4.571	0.465	0.48
0.456	0.504	0.48	0.518	0.643	4.411	0.48	0.48
0.705	0.536	0.78	0.518	0.675	4.445	0.78	0.65
0.738	0.905	0.65	0.508	0.618	3.74	0.65	0.78
7.839	8.496	9.175	7.55	6.781	4.701	9.175	6.5
9.04	9.483	9.138	9.12	7.019	5.591	9.25	7.25
8.221	8.905	9.25	8.75	6.928	5.34	9.25	9.1
8.001	7.85	9.1	6.576	6.576	4.142	9.1	9.25
8.86	8.496	7.25	7.55	7.147	4.837	7.25	9.25
10.637	7.85	7.25	6.576	5.555	4.546	7.25	11.5

Table 7: Sample of Selling price (in hundred thousand) and different machine learning using Federated Learning models predicted selling price values.

Fed-Linear Regression	Fed-KNN	Fed-AdaBoost	Fed-Random Forest	Fed-Neural Network	Fed-Gradient Boosting	Fed- SVM	Selling Price
0.365	0.508	0.48	0.491	0.365	0.466	0.74	0.4
0.409	0.558	0.48	0.557	0.409	0.402	0.751	0.45
0.409	0.558	0.48	0.557	0.409	0.402	0.751	0.48
0.437	0.492	0.45	0.451	0.436	0.444	0.698	0.48
0.423	0.492	0.48	0.537	0.423	0.48	0.761	0.48
0.676	0.796	0.78	0.75	0.676	0.78	0.748	0.65
0.661	0.622	0.65	0.673	0.661	0.65	0.815	0.78
7.78	8.12	9.175	9.007	7.78	9.175	6.926	6.5
8.803	8.15	9.1	8.193	8.803	9.1	6.438	7.25
8.268	8.03	9.25	8.721	8.268	9.245	7.044	9.1
8.399	8.15	9.1	8.193	8.399	9.1	6.462	9.25
8.785	8.15	9.1	8.193	8.785	9.1	6.44	9.25
11.078	8.27	7.25	7.693	11.078	7.253	5.716	11.5

Table 5 presents a comparison between the accuracy of machine learning models on the automobile dataset using a federated learning strategy. Table 5 presents a summary of the findings of the algorithms. The major distinction between the algorithms after using federated learning is that Linear Regression achieves the better results, based on the observed results in this table.

5.3 Dickey-Fuller test

The Dickey-Fuller test is an important statistical test in time series analysis. It is used to evaluate whether a time series is stationary or not. Stationary is an essential property of time series data because it allows for consistent and predictable patterns to emerge[33], [34], [35].

In time series analysis, correct forecasting requires that the data be stationary. If the data is not stationary, the forecasts may be inaccurate or unstable. By converting a non-stationary time series to a stationary one using the Dickey-Fuller test, it will improve the accuracy of forecasting models. Furthermore, the Dickey-Fuller test is significant in statistical modeling as it helps in showing the order of integration for a given time series. It aids in deciding if data needs to be different or transformed to make it proper for modeling. The test statistic for the Dickey-Fuller test is calculated as:

$$d_t = (y_t - y_{t-1}) - d_{t-1} \qquad (13)$$

Where y_t is the value of the time series at time t, d_t is the differenced series at time t, and d_{t-1} is the value of the differenced series at time t-1. The null hypothesis of the Dickey- Fuller test is that the time series is non-stationary (i.e., has a unit root). The alternative hypothesis is that the time series is stationary (i.e., does not have a unit root).

The critical values of the test statistic are decided based on the sample size and the level of significance chosen for the test. If the test statistic is less than the critical value, then the null hypothesis is rejected, and the time series is stationary. If the test statistic is greater than the critical value, then the null hypothesis is not rejected, and the time series is non-stationary. In summary, the Dickey-Fuller test is important in converting time series to accuracy because it helps to check for trend and seasonality in the data, finds non-stationary in a time series, and aids in selecting the proper transformation for correct modeling and forecasting.

Table 8: Sample of machine learning models accuracy after applying federated learning.

Model Name	Model Accuracy	Model training Time	Model Name	Model Accuracy	Model training Time
Fine Tree	0.67	2.12	**Fed-Fine Tree**	0.68	0.40
Medium Tree	0.70	0.56	**Fed-Medium Tree**	0.67	0.35
Course Tree	0.78	0.42	**Fed-Course Tree**	0.79	0.37
Linear Discriminant	0.76	0.86	**Fed-Linear Discriminant**	0.75	0.35
Linear SVM	0.78	3.26	**Fed-Linear SVM**	0.78	2.58
Quadratic SVM	0.72	1.41	**Fed-Quadratic SVM**	0.72	1.29
Cubic SVM	0.68	1.34	**Fed-Cubic SVM**	0.69	1.23
Fine Gaussian SVM	0.78	0.47	**Fed-Fine Gaussian SVM**	0.78	0.37
Medium GaussianSVM	0.78	0.44	**Fed-Medium GaussianSVM**	0.78	0.36
Course GaussianSVM	0.78	0.41	**Fed-Course GaussianSVM**	0.78	0.43
Fine KNN	0.68	0.69	**Fed-Fine KNN**	0.69	0.37
Medium KNN	0.78	0.40	**Fed-Medium KNN**	0.78	0.32
Course	0.78	0.41	**Fed-**	0.78	0.31

KNN			**Course KNN**		
Cosine KNN	0.78	0.46	**Fed-Cosine KNN**	0.78	0.31
Cubic KNN	0.78	0.52	**Fed-Cubic KNN**	0.78	0.35
Weighted KNN	0.76	0.40	**Fed-Weighted KNN**	0.75	0.31
Ensemble Boosted Tree	0.75	11.36	**Fed-Ensemble Boosted Tree**	0.74	9.60
Ensemble Subspace Discriminant	0.78	8.46	**Fed-Ensemble Subspace Discriminant**	0.78	7.74
Ensemble Subspace KNN	0.70	8.92	**Fed-Ensemble Subspace KNN**	0.72	8.20
Ensemble RUS Boosted Trees	0.58	10.01	**Fed-Ensemble RUS Boosted Trees**	0.55	9.62
ARIMA	0.68	0.027	**Fed-ARIMA**	0.69	0.022
VAR	0.7	0.037	**Fed-VAR**	0.72	0.017

Figures 7 and figure 8 show a comparison between the used machine learning models before and after using the federated learning model. The result of the proposed model in MATLAB 20 applied upon 4 cores in GPU only using a cars inventory database.

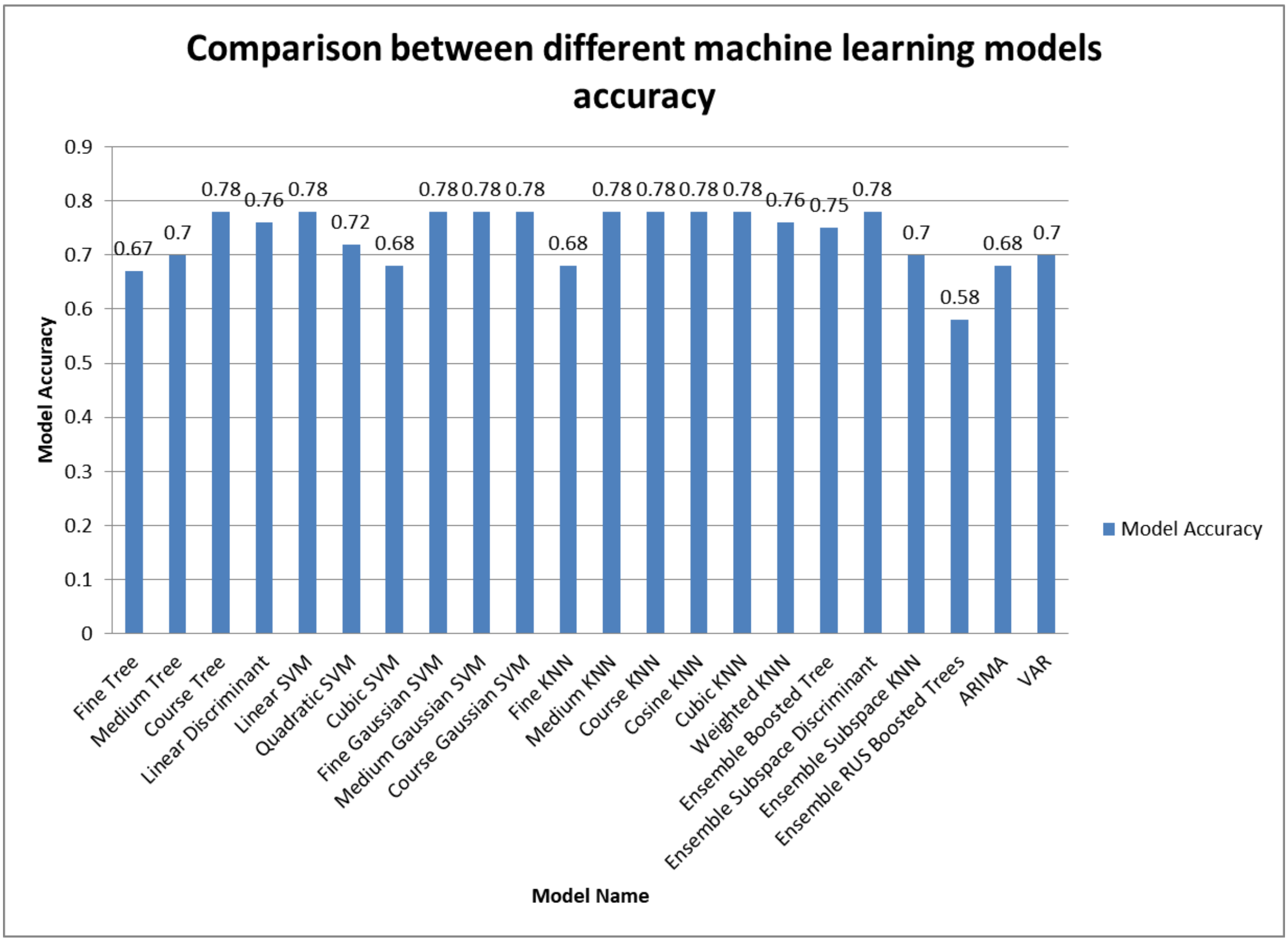


Fig 7: Comparison between different machine learning models accuracy.

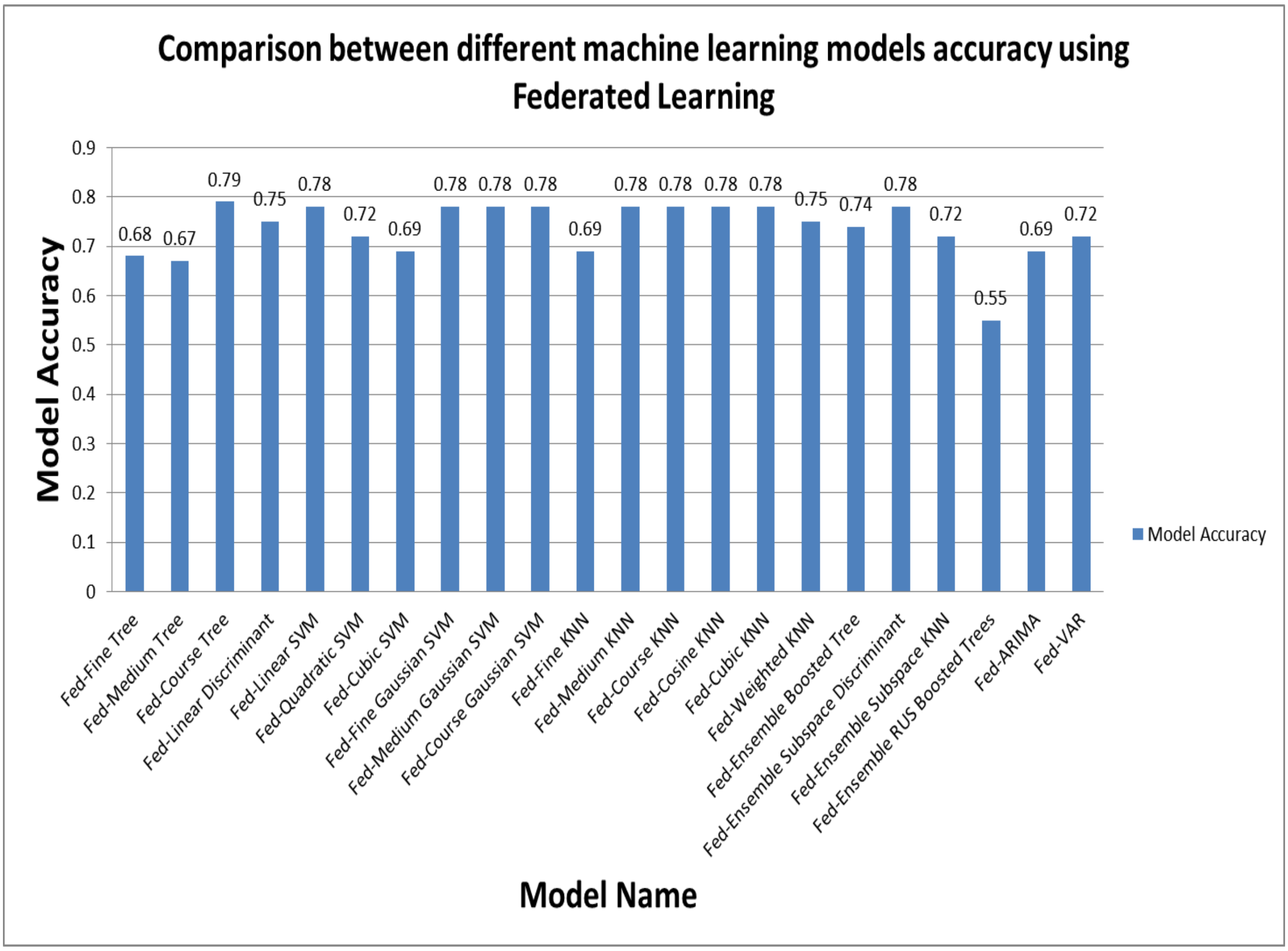


Fig 8: Comparison between different machine learning models accuracy using Federated Learning

6. CONCLUSION AND FUTURE DIRECTION

Internet word-of-mouth in online platforms and forums can now directly affect consumers' buying decisions thanks to the Internet's and artificial intelligence's recent quick development. As a result, this research suggests a multi-factor model for predicting automotive sales.

The Federated Learning benefits from robust generalization, simple structure, and global optimization. The fact that it is proper for multidimensional and small sample data—which are consistent with the traits of influencing variables and sales data—is what matters most. As a result, model training uses the Fed-Learning model. It is used to intelligently perfect the model parameters to increase model accuracy and avoid the impact of the kernel function and penalty factor on model performance.

For the trials, the Kaggle dataset, and for the model's performance evaluation are used, the MAPE and RMSE values were used. First, this study summarizes the elements affecting the sales of automobiles. It finds that there are six external influencing factors: steel output, rubber tire output, money supply, diesel output, consumer price index, and vehicle pricing. The MAPE and RMSE, after incorporating the Kaggle data into the model, are 0.35 and 0.29, respectively.

This work develops the Federated Learning automotive sales forecast model based on criteria takings into consideration the star ratings of users in online forums, which is of considerable guiding significance for the automobile industry, in view of the prediction problem of monthly automobile sales. Still, there are certain gaps in the findings that require further investigation in the future.

(1) Consumers' online review data includes significant emotional information about them. The (feature, emotional value) pairings in the review data can be mined using the sentiment analysis method to boost prediction accuracy.

(2) The deep learning algorithm's supremacy in the prediction problem has been established with the rise of this technique. So, to increase forecast efficiency and accuracy, it can be thought about using the deep learning algorithm for the automotive sales problem.

Additionally, investigation of the variables that influence auto sales and observe how the influencing variables are related.

7. ACKNOWLEDGMENTS

I would like to express the utmost gratitude to my supervisor, Mohamed Belal, for supplying valuable guidance and support. His knowledge and ability in the field have been invaluable in shaping the direction of the research. I would also like to thank the participants who generously gave their time and insights to the study. Without their contributions, this research would not have been possible.

Additionally, I would like to acknowledge Sara Mohammed for her aid with data analysis, literature review, and manuscript preparation. Finally, thankful to families and loved ones for their unwavering support and encouragement throughout this journey. Thank you all for your contributions to this research.

Ethics declarations

Conflict of interest / Declarations

The authors declare that there is no conflict of interest in the publication of this paper.

Competing Interests

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for- profit sectors. No funding was received to aid with the preparation of this manuscript.

Data Availability Statements

The four real-world data sets are publicly available from the links provided in the paper. The synthetic data sets are available from the corresponding author in the following repository: https://github.com/MohammedMahfouz/InventoryDataset.git

Declarations / Funding

No funding was received to aid with the preparation of this manuscript.

Author contributions

The authors confirm contribution to the paper as follows: study conception and design: Mohammed A. Mahfouz1 and Sara M. Mosaad2; data collection: Mohammed A. Mahfouz1; analysis and interpretation of results: Mohammed A. Mahfouz1, Sara M. Mosaad2 and Mohammed A. Belal3; draft manuscript preparation: Mohammed A. Mahfouz1. All authors reviewed the results and approved the last version of the manuscript.

Corresponding author

Correspondence to Mohammed A. Mahfouz.

8. REFERENCES

[1] B. V. R. Reddy and Dr. K. S. Sree, "Car Price Prediction Using Machine Learning Algorithms," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 8, pp. 1093–1099, Aug. 2022, doi: 10.22214/ijraset.2022.46354.

[2] F. Qu *et al.*, "Forecasting of Automobile Sales Based on Support Vector Regression Optimized by the Grey Wolf Optimizer Algorithm," *Mathematics*, vol. 10, no. 13, p. 2234, Jun. 2022, doi: 10.3390/math10132234.

[3] T. Kostyra and M. Rubaszek, "Forecasting the Yield Curve for Poland," *Econom. Res. Finance*, vol. 5, pp. 103–117, Dec. 2020, doi: 10.2478/erfin-2020-0006.

[4] G. Kulkarni, P. K. Kannan, and W. Moe, "Using online search data to forecast new product sales," *Decis. Support Syst.*, vol. 52, no. 3, pp. 604–611, Feb. 2012, doi: 10.1016/j.dss.2011.10.017.

[5] X. Li, C. Wu, and F. Mai, "The effect of online reviews on product sales: A joint sentiment-topic analysis," *Inf. Manage.*, vol. 56, no. 2, pp. 172–184, Mar. 2019, doi: 10.1016/j.im.2018.04.007.

[6] Y.-Y. Wang, T. Wang, and R. Calantone, "The effect of competitive actions and social media perceptions on offline car sales after automobile recalls," *Int. J. Inf. Manag.*, vol. 56, p. 102257, Feb. 2021, doi: 10.1016/j.ijinfomgt.2020.102257.

[7] H. Hu, L. Tang, S. Zhang, and H. Wang, "Predicting the direction of stock markets using optimized neural networks with Google Trends," *Neurocomputing*, vol. 285, pp. 188–195, Apr. 2018, doi: 10.1016/j.neucom.2018.01.038.

[8] X. Li, P. Wu, and W. Wang, "Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong," *Inf. Process. Manag.*, vol. 57, no. 5, p. 102212, Sep. 2020, doi: 10.1016/j.ipm.2020.102212.

[9] K. Liu, J. Zhou, and D. Dong, "Improving stock price prediction using the long short-term memory model combined with online social networks," *J. Behav. Exp. Finance*, vol. 30, p. 100507, Jun. 2021, doi: 10.1016/j.jbef.2021.100507.

[10] S. Prasanth, U. Singh, A. Kumar, V. A. Tikkiwal, and P. H. J. Chong, "Forecasting spread of COVID-19 using google trends: A hybrid GWO-deep learning approach," *Chaos Solitons Fractals*, vol. 142, p. 110336, Jan. 2021, doi: 10.1016/j.chaos.2020.110336.

[11] "A deep learning approach on short-term spatiotemporal distribution forecasting of dockless bike-sharing system | SpringerLink." https://link.springer.com/article/10.1007/s00521-018-3470-9 (accessed Mar. 10, 2023).

[12] Z.-P. Fan, Y.-J. Che, and Z.-Y. Chen, "Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis," *J. Bus. Res.*, vol. 74, pp. 90–100, May 2017, doi: 10.1016/j.jbusres.2017.01.010.

[13] K. Kim and H. Ahn, "A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach," *Comput. Oper. Res.*, vol. 39, no. 8, pp. 1800–1811, Aug. 2012, doi: 10.1016/j.cor.2011.06.023.

[14] "(PDF) Forecasting Tourism Demand with Google Trends for a Major European City Destination." https://www.researchgate.net/publication/282751602_Forecasting_Tourism_Demand_with_Google_Trends_For_a_Major_European_City_Destination (accessed Mar. 10, 2023).

[15] F. Moussa, E. Delhoumi, and O. B. Ouda, "Stock return and volatility reactions to information demand and supply," *Res. Int. Bus. Finance*, vol. 39, pp. 54–67, Jan. 2017, doi: 10.1016/j.ribaf.2016.07.016.

[16] S. Shakti, M. Hassan, Y. Zhenning, R. Caytiles, and N. C. S. N. Iyenger, "Annual Automobile Sales Prediction Using ARIMA Model," *Int. J. Hybrid Inf. Technol.*, vol. 10, pp. 13–22, Jun. 2017, doi: 10.14257/ijhit.2017.10.6.02.

[17] A. Sa-ngasoongsong, S. T. S. Bukkapatnam, J. Kim, P. S. Iyer, and R. P. Suresh, "Multi-step sales forecasting in automotive industry based on structural relationship identification," *Int. J. Prod. Econ.*, vol. 140, no. 2, pp. 875–887, Dec. 2012, doi: 10.1016/j.ijpe.2012.07.009.

[18] K. N. Konstantakis, C. Milioti, and P. G. Michaelides, "Modeling the dynamic response of automobile sales in troubled times: A real-time Vector Autoregressive analysis with causality testing for Greece," *Transp. Policy*, vol. 59, pp. 75–81, Oct. 2017, doi: 10.1016/j.tranpol.2017.07.006.

[19] S. Ding and R. Li, "Forecasting the sales and stock of electric vehicles using a novel self-adaptive optimized grey model," *Eng. Appl. Artif. Intell.*, vol. 100, p. 104148, Apr. 2021, doi: 10.1016/j. engappai.2020.104148.

[20] L.-Y. He, L.-L. Pei, and Y.-H. Yang, "An optimised grey buffer operator for forecasting the production and sales of new energy vehicles in China," *Sci. Total Environ.*, vol. 704, p. 135321, Feb. 2020, doi: 10.1016/j.scitotenv.2019.135321.

[21] S. Ding, R. Li, and S. Wu, "A novel composite forecasting framework by adaptive data preprocessing and optimized nonlinear grey Bernoulli model for new energy vehicles sales," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 99, p. 105847, Aug. 2021, doi: 10.1016/j.cnsns.2021.105847.

[22] W. Chen, H. Zhang, M. K. Mehlawat, and L. Jia,

"Mean–variance portfolio optimization using machine learning-based stock price prediction," *Appl. Soft Comput.*, vol. 100, p. 106943, Mar. 2021, doi: 10.1016/j.asoc.2020.106943.

[23] M. Vijh, D. Chandola, V. A. Tikkiwal, and A. Kumar, "Stock Closing Price Prediction using Machine Learning Techniques," *Procedia Comput. Sci.*, vol. 167, pp. 599–606, Jan. 2020, doi: 10.1016/j.procs.2020.03.326.

[24] Y. Peng and W. Xiang, "Short-term traffic volume prediction using GA-BP based on wavelet denoising and phase space reconstruction," *Phys. Stat. Mech. Its Appl.*, vol. 549, p. 123913, Jul. 2020, doi: 10.1016/j.physa.2019.123913.

[25] H. Yang, X. Li, W. Qiang, Y. Zhao, W. Zhang, and C. Tang, "A network traffic forecasting method based on SA optimized ARIMA–BP neural network," *Comput. Netw.*, vol. 193, p. 108102, Jul. 2021, doi: 10.1016/j.comnet.2021.108102.

[26] "1985 Automobile Dataset." https://www.kaggle.com/datasets/fazilbtopal/auto85 (accessed Mar. 05, 2023).

[27] "Automobile Dataset." https://www.kaggle.com/datasets/toramky/automobile-dataset (accessed Mar. 05, 2023).

[28] "AutoMobile Dataset." https://www.kaggle.com/datasets/mrushan3/automobile-dataset (accessed Mar. 05, 2023).

[29] "Pakistan's Largest Automobile Listing - PakWheels." https://www.kaggle.com/datasets/muhammadwaqargul/pakwheels-used-car-dataset-october-2022 (accessed Mar. 05, 2023).

[30] D. Chicco, M. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, p. e623, Jul. 2021, doi: 10.7717/peerj-cs.623.

[31] D. Kwiatkowski and P. Schmidt, "Dickey-fuller tests with trend," *Commun. Stat. - Theory Methods*, vol. 19, no. 10, pp. 3645–3656, Jan. 1990, doi: 10.1080/03610929008830402.

[32] J. Otero and C. F. Baum, "Unit-root Tests Based on Forward and Reverse Dickey–Fuller Regressions," *Stata J.*, vol. 18, no. 1, pp. 22–28, Mar. 2018, doi: 10.1177/1536867X1801800103.

[33] R. Mushtaq, "Augmented Dickey Fuller Test," *SSRN Electron. J.*, Aug. 2011, doi: 10.2139/ssrn.1911068.

[34] I. E. Mohamed, "Time Series Analysis Using SAS - Part I - The Augmented Dickey Fuller (ADF) Test," 2008.

[35] F. X. Diebold and G. D. Rudebusch, "On the power of Dickey-Fuller tests against fractional alternatives," *Econ. Lett.*, vol. 35, no. 2, pp. 155–160, Feb. 1991, doi: 10.1016/0165-1765(91)90163-F.