Gusau Journal of AccountingandFinance
(GUJAF)

Vol.5Issue1,April,2024ISSN:2756-665X

A Publication of
DepartmentofAccountingandFinance,
Faculty of Management and Social Sciences,
FederalUniversityGusau,ZamfaraState-Nigeria

Vol.5Issue1
April, 2024
ISSN:2756-665X

A Publication of
DepartmentofAccountingandFinance,
Faculty of Management and Social Sciences,
FederalUniversityGusau,ZamfaraState-Nigeria

Publishedandprinted by:
AhmaduBelloUniversityPressLimited,Zaria Kaduna State, Nigeria.
Tel:08065949711,069-879121
e-mail:
abupress2013@gmail.comabupress2020@yahoo.com
Website:www.abupress.com.ng

EDITORIALBOARD

CALLFOR PAPERS

The editorial board of Gusau Journal of Accounting and Finance (GUJAF) is hereby inviting authors to submit their unpublished manuscript for publication. The journal is published in two issues of April and October annually. GUJAF is a double-blind peer reviewed journal published by the Department of Accounting and Finance, Faculty of Management and Social Sciences, Federal UniversityGusau, Zamfara State Nigeria The Journal accepts papersin all areas of Accounting and Finance for publication which include: Accounting Standards, Accounting Information System, Financial Reporting, Earnings Management, , Auditing and Investigation, Auditing and Standards, Public Sector Accounting and Auditing, Taxation and Revenue Administration, Corporate Governance Issues, Corporate Social Responsibility, Sustainability and Environmental Reporting Issue, Information and Communication TechnologyIssues, BankruptcyPrediction, Corporate Finance, Personal Finance, Merger and Acquisitions, Capital Structure, Working Capital Management, Enterprises Risk Management, Entrepreneurship, International Business Accounting and Finance, Banking Crises, Bank‘s Profitability, Risk and Insurance Issue, Islamic Finance, Conventional and Islamic Banks and so forth.

GUIDELINESFORSUBMISSIONANDMANUSCRIPT FORMAT

The submission language is English and must be a well-researched original manuscript that has not previously been submitted elsewhere for publication. The paper should not exceed more than 15 pages on A4 type paper in MS-word format, 1.5-line spacing, 12 Font size in Times new roman. Manuscript should be tested for plagiarism before submission, as the maximum similarity index acceptable by GUJAF is 25 percent. Furthermore, the length of a complete article should not exceed 5000 words including an abstract of not more than 250 words with a minimum of four key words immediately after the abstract. All references including in text citation and reference list, tables and figures should be in line with APA 7th Edition publication manual. Finally, manuscript should be send to our email address elfarouk105@gmail.comand a copy to our website on journals.gujaf.com.ng

PUBLICATION PROCEDURE

After receiving a manuscript that is within the similarity index threshold, a confirmation email will be send together with a request to pay a review proceeding fee.At this point, the editorial board will take a decision on accepting, rejecting or making a resubmission of the manuscript based on the outcome of the double-blind peer review. Those authors whose manuscript were accepted for publication will be asked to pay a publication fee, after effecting all suggested corrections and changes made on the manuscript. All corrected papers returned within the specified time frame will be published in that issue.

PAYMENTDETAILS
Bank:FCMB
AccountNumber:7278465011
AccountName:Gusau Journalof Accountingand Finance

FORINQUIRY
TheHead,
Department of Accounting and Finance,
FederalUniversityGusau,ZamfaraState.elfarouk
105@gmail.com
+2348069393824

FORMOREINFORMATION, CONTACT
TheEditor-in-Chiefon+2348067766435
TheAssociateEditoron+2348036057525
ORvisit ourwebsiteonwww.gujaf.com.ngorjournals.gujaf.com.ng

FORECASTING AUTOMOBILE DEMANDANDSALESINTHENIGERIANMARKET: A MACHINE LEARNING APPROACH TOURBAN MOBILITY, MARKET COMPETITION, AND POLICY INSIGHTS

Emmanuel Imuede Oyasor
Department of Accounting Science,
Walter Sisulu University, Mthatha, SouthAfrica
emmanueloyasor247@gmail.comhttps://doi.org/10.57233/gujaf.v5i1.20

Abstract
Theestimationofautomobiledemandiscentraltobothacademic inquiryandpolicyplanning,particularlygiven the sector‘s critical role in global economic activity. In developed economies such as the United States, Germany, and China, the auto industry serves as a paradigmatic case for analyzing market dynamics in differentiated, oligopolistic settings. Accurate demand forecasting is essential for production planning, pricing strategy, and infrastructure development. However, in emerging markets like Nigeria, empirical research on automobile demand remains sparse despite its growing relevance. Nigeria's automotive landscape is undergoing rapid transformation, propelled by urbanization, a rising middle class, and industrial policy reforms such as the National Automotive Industry Development Plan (NAIDP). This study addresses the empirical gap byevaluating the performance of various regression models, including the OLS, MARS, Regression Tree, Random Forest, and Gradient Boosting, in predicting automobile demand using real-world data. Among the models tested, OLS emerged as the most effective, with the lowest error metrics (MAE = 0.15, MSE = 0.06, RMSE = 0.24) and a strong explanatory power (R^2 = 0.86). In contrast, the MARS model underperformed, displaying the highest error rates and limited predictive capacity (R^2 = 0.43). Ensemble methods (RF and GB) showed moderate performance, with GB slightly outperforming RF in terms of relative error (MAPE = 0.01). The RegressionTree modelalsoperformed well,balancingaccuracyand interpretability.Thefindingsoffer valuable insights for both policymakers and industry stakeholders in Nigeria, emphasizing the importance of model selection in automotive demand estimation and the strategic implications for infrastructure and investment planning.

Keywords: Automobiledemand,demandforecasting,regressionmodels,machinelearning,policyplanning
JELCodes:C53,L62,R41,O55

1.0 Introduction

The estimation of automobile demand has remained a prominent topic in both academic and policy discourse for decades. As a key sector in the global economy, generating significant revenues and employment in countries such as the United States, Germany, and China, the automobile industry has attracted widespread scholarly attention. Accurate demandestimation is crucial for firm-level decision-making, particularly in forecasting production capacity and projecting sales revenues. For economists, the auto industry represents a textbook example of an oligopolistic, differentiated products market, making it ideal for studying pricing strategies, market power, and demand elasticity. Likewise, policymakersfind automobile demand estimation indispensable for planning infrastructure needs, such as road expansion and maintenance.

In Nigeria, the automotive market is rapidly evolving, driven by a growing middle class, increasing urbanization, and government policies promoting vehicle assembly and electric mobility. Reliable empirical insights into automobile demand remain limited, despite the market‘s complexity and strategic importance. As Nigeria continues to develop its transport infrastructure and vehicle assembly capabilities, especially under initiatives like the National Automotive Industry Development Plan (NAIDP), accurate demand forecasting becomes even more essential for public planning and private investment.

Literature has amassed extensive reviews focused on keytransport planning parameters, such as demand forecasting and appraisal, largely spurred by the data collection capabilities of researchers.Althoughnotallofthesestudiescontributenewprimarydata,theyarefrequently cited and widely regarded as valuable contributions to the transport planning canon. There is also growing empirical evidence supporting the forecasting of automobile prices and sales volumes. In market-driven economies, accurate sales forecasts are fundamental to strategic planning and operational efficiency. To this end, a variety of forecasting models have been adopted.One ofthemost notableis the Bass diffusion model, renowned forits simplicityand predictive accuracy. This model explains how new products penetrate markets by capturing the dynamics between innovators and imitators. It has been widelyapplied in sectors ranging from consumer goods to digital platforms (Øverby et al., 2023; Han & Tang, 2022; Zhang et al., 2022).

However, as markets become increasingly volatile and influenced by vast streams of real-time data, classical models often struggle to maintain accuracy. In Nigeria‘s case, where informal markets, policy shifts, and economic fluctuations introduce added layers of uncertainty, the limitations of traditional forecasting techniques become particularly pronounced. The integration of machine learning (ML) models is gaining momentum. These models leverage statistical learning from historical datasets to enhance prediction accuracy. Bao et al. (2022) introduced a Relational Vector Machine (RVM) approach for small-sample regression, demonstrating its relevance to predicting electric vehicle (EV) ownership. Yet, such models may still fall short in addressing data volatility and multidimensional indicator complexity, especially in the Nigerian context where data granularity and consistency can be limited.

To overcome these challenges, deep learning (DL) has emerged as a more robust alternative. Unlike traditional ML techniques, DL models excel at identifying complex, non-linear relationships and adapting to dynamic data environments. They have been successfully deployed in fields such as car ownership forecasting, energy consumption modeling, and carbon emissions estimation (Qiao et al., 2021). DL models are particularly advantageous in settings like Nigeria, where small sample sizes and inconsistent data reporting hinder conventional forecasting methods. These models possess high self-learning capabilities, enabling them to derive meaningful insights even from noisy or incomplete datasets (Feng& Chen, 2021).

One of the primary strengths of DL is its ability to reduce errors stemming from data redundancy and random noise through sophisticated input functions (Zhu et al., 2019). Additionally, model robustness in small-sample environments can be significantly improved through data augmentation strategies, which increase both the volume and variability of training data (Zeng et al., 2017; Hong et al., 2022). A notable example is Liu et al. (2021), whoenhanceddatadiversityusingDiscreteWaveletTransformation(DWT)combinedwitha hybrid Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) model, effectively capturing complex non-linear data patterns.

The application of advanced forecasting techniques such as ML and DL offers a promising path for improving automobile demand estimation. The methods can enhance policy formulation, business strategy, and infrastructure planning in an increasingly complex and data-scarce marketenvironment.Theremainderof thispaperisstructuredasfollows:Section

2 presents a comprehensive review of the empirical literature; Section 3 outlines the data, model specifications, and estimation strategies; Section 4 reports and discusses the empirical results, and policy implications. Section 5 concludes with recommendations, limitations, and future research directions.

2.0 LiteratureReview

Sales forecasting is crucial for demand-driven supplychains as it helps companies efficiently manage production, inventory, resource, and services (Sohrabpour et al., 2021). Accurate sales forecasting assists businesses in formulating more reasonable short-term operational plans, as well as medium to long-term plans at the tactical and strategic levels (Gustriansyah et al., 2022). One of the challenges faced in sales forecasting is the complex nonlinear characteristics of sales data under the influence of internal or external multi-source factors. The accuracy of sales prediction is greatly influenced using relevant features. In sales forecasting research, some methods are based on univariate prediction using sales history data, such as simple moving average, exponential smoothing method and its variants (Croston, 1972)- (Yang et al., 2021), ARIMA and its variants (Rostami-Tabar et al., 2023)- (Londhe and Palwe, 2022), and state space models (Svetunkov and Boylan, 2023); (de Rezende et al., 2022). However, it is difficult to generate accurate forecasts solely relying on historical observations (Sareminia and Amini, 2023). The lack of relevant features mayresult in fast-changing patterns in the sales series being considered as noise, to deteriorated model performance.

Another school of researchers utilize multivariate prediction methods to achieve more accurate forecasts. The features used by them include weather, calendar attributes, and promotions, as well as lagged variables. For example, Di Pillo et al. (Di Pillo et al., 2016) employed support vector machine to predict the daily sales volume of a certain type of pasta under aperiodic promotional events, using 13-dimensional features including calendar attributes and product dimensions. Weng et al. (Weng et al., 2019) achieved outstanding results by constructing time series features, statistical features, and detail features on a publicly available sales dataset. Pan and Zhou (Pan and Zhou, 2020) conducted sales forecasting by using online sales features such as product, price, search, and browsing as training features, utilizing convolutional neural network (CNN) on an e-commerce dataset.He et al. (He et al., 2022) implemented multivariate forecasting based on LSTM and particle swarm optimization algorithm on three classic sales datasets.

Andrade et al. (Andrade and Cunha, 2023) used extreme gradient boosting (XGBoost) to account for sales fluctuations caused by external factors and achieved retail forecasting. According to previous research, multivariate prediction methods with exogenous variables exhibit higher accuracy compared to univariate prediction methods (Fildes et al., 2022). However, these exogenous variables may contain both time dependent and time-independent components. Existing literature has paid little attention to the application of different processing methods for different features, and effective methods for extracting multiple features have not been found. Therefore, one of the issues that this paper aims to address is how to effectively select features and extract accurate information from them.

Moreover,prediction based on data-driven approaches often faces the issueofoverfittingdue to insufficient data samples, especially in sales forecasting for new products. To address this, some scholars have utilized transfer learning methods to enhance data availability, so as to improvethepredictiveaccuracy(Lyuetal.,2023).Forinstance,Afrinetal.(Afrinetal.,

2018) achieved earlydemand forecasting for new products byutilizing historical information for existing products and a demand differentiation index between new and existing products. Fan et al. (Fan et al., 2023) improved the prediction accuracy of multiple components in the equipment aftermarket through transfer learning by performing similar clustering and joint representation learning.

Schneider and Gupta (Schneider and Gupta, 2016) implemented sales forecasting for new productsusingsaleshistorydataofexistingproducts.Thesestudieshaveprovidedinspiration for us to enhance the predictive accuracy of models using data from similar products. Intermediate level series. Subsequently, higher and lower level forecasts are obtained by aggregation and disaggregation of the MO forecasts (Karmy and Maldonado, 2019).Although these three methods satisfy the consistency of hierarchical forecasting, they often introduce biases during the aggregation or disaggregation process, leading to inadequate prediction accuracy. Indeed, it is possible to independently predict all series, disregarding their interdependence. However, the resulting predictions are unlikely to satisfy consistency and may have significant deviations from reality

Based on uncertain environmental scenarios such as COVID-19, Ma et al. (2019) predicted EV sales in 20 countries by the Bass diffusion model. Under the role of different technological advances, economic development, and policy incentives, Rietmann et al.(2020)made along-term forecastofEVownershipin26countriesonfive continentsthrough a Logistic model, and the global EV market penetration will reach 30% by 2032. Sun and Wang et al. (2022) develop a system dynamics (SD) model of China‘s EV market evolution based on the competitive Lotka-Volterra (LV) model, where EVs will gradually replace fuel vehicles and dominate the vehicle market by 2050. The grey model has been widely used in forecasting the EVs sales and ownership due to its good applicability for small sample forecasting (Ding and Li, 2021). The traditional grey model can be further optimized by a grey buffer operator with a genetic algorithm (He et al., 2020), by fitting the nonlinear relationship between the grey information factor and the time factor (Liu et al., 2022).

Machine learning uses statistical models learned on pre-prepared training samples to achieve accurate prediction. Bao et al. (2022) implemented a relational vector machine (RVM) approach to mine regression relationships from available data, which shows some applicability to the problem of EV ownership in small samples. The above research methods failed to effectively eliminate the volatility of data, which failed to achieve EV sales long-term accurate prediction with multiple research objects and multiple indicator dimensions.

Liu B (2023) Policy incentives are the key driving force for the electric vehicle (EV) market cultivation. During the EV market cultivation in different regions, the supply-side and demand-side policies have different effects. Accurately forecasting the stage characteristics and response sensitivity of EV sales under supply-demand side policy scenarios, which is crucial to the EV promotion policies design. This study selects the EV sales in 31 provinces with data available, as the basis for decision-making, and proposes a multi-factor prediction model integrating grey relation analysis (GRA), discrete wavelet transform (DWT), and bidirectional long short-term memory. Combined with the development difference of 31 provinces, the penetration of China‘s EV market under the benchmark, supply, demand, and ideal scenarios are verified. The experimental results show that the average Mean Absolute Percentage Error (MAPE) of the GRA-DWT-BiLSTM model is 9.884, and the 31 samples showgoodapplicabilityforEVsalesforecasting.In2027,thegrowthrateofChina‘sEV sales in the demand-side scenario will exceed the supply-side scenario. Under the ideal scenario,China‘sEVpenetrationratewillreach 27.31%,42.40%,and 52.97%in2024,2030, and 2035 respectively. The forecast results provide a decision-making basis for China‘s EV market sequential supply-demand side policies

3.0 Methodology

This study adopts a comparative predictive modelling framework to estimate automobile prices in Nigeria using both classical statistical and advanced machine learning algorithms. Thegoal is toevaluateand compare model performancein terms oftheiraccuracyand ability to generalize across multiple dimensions of automobile characteristics such as mileage, engine horsepower, cylinder count, and observed sale prices.

The dataset comprises 5,000+ observations of vehicles sold in the Nigerian market, capturing key variables such as Price, Mileage, Cylinders, and Horsepower (Hp). Table 1 provides the descriptivestatisticsofthedataset.Theaveragepriceofvehiclesisapproximately ₦4,514,644 with a standard deviation of ₦5,500,000, indicating significant variability in the market, likely due to the presence of both luxury and economy cars. Similarly, mileage and horsepower vary widely, reflecting different usage patterns and vehicle types.

Table 1:
Descriptivestatistics ofthe data

Variable	Mean	Median	Std	Min	Max
Price	4514644		4297011	550000	62400000
Milage	194984	176290	139576	1	2456318
Cylinder	5.16	6	1	4	8
Hp	208.83	203	70	83	585

Source: Author (2024)

Themethodological framework employed fivemodels for predictingvehicleprices: Ordinary Least Squares (OLS) Regression, Multivariate Adaptive Regression Splines (MARS), RegressionTree(CART),RandomForest(RF)andGradientBoosting(GB).Eachmodelwas trained and evaluated on an 80-20 train-test split of the dataset. Performance metricsincluding Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean SquareError (RMSE), Mean Absolute Percentage Error (MAPE), R^2 Score, and Explained Variance Score (EVS) were computed for both training and testing phases.

Themathematicalformulationofthepredictiontaskisspecifiedasfollows: Let $y_i$bethelog-transformed vehicle price, and $x_i = [x_{i1}, x_{i2}, x_{i3}, x_{i4}]$ denote the feature vector for each observation representing mileage, cylinders, horsepower, and other attributes. The prediction function can be denoted as:

$$\hat{y}_i = (x_i) \tag{1}$$

Where$(\cdot)$ istheestimatedmodelfunctionlearnedviathetrainingalgorithm. The OLS model assumes a linear relationship between predictors and price:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \tag{2}$$

The Random Forest model, by contrast, constructs multiple decision trees and averages theoutputs:

$$\hat{y}_i^{RF}=\frac{1}{T}\sum_{t=1}^{T} h(x)_i \quad (3)$$

Where $h(\cdot)$ is the output of the t-th tree in the ensemble.

GradientBoostingbuildstreessequentiallytominimizeprediction error:

$$F(x)=F_{m-1}(x)+\gamma_m h_m(x) \quad (4)$$

Where $h(x)$ isthenewweaklearnerfittedtotheresidualsofthepreviousmodel.

The models were assessed on both the training and testing datasets. Accuracy, precision, recall, F1-score, specificity, Matthews Correlation Coefficient (MCC), and Cohen‘s Kappa were also calculated, particularly for models adapted for classification sub-tasks (e.g., price category prediction).

The R^2 scoremeasurestheproportionofvarianceinthedependentvariable explainedbythe model:

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \quad (5)$$

TheMAEandRMSEprovide absoluteandsquareddeviations respectively:

$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \quad (6)$$

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}(y_i-\hat{y}_i)^2} \quad (7)$$

TheMAPEmetricquantifiesrelativeprediction error:

$$\text{MAPE}=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \quad (8)$$

4.0 ResultandPolicyImplications

The performance of various regression models was evaluated using standard statistical metrics. The OLS model attained the lowest MAE (0.15), MSE (0.06), and RMSE (0.24), indicating that it had the smallest average error and variance in predictions compared to the other models. Additionally, its R^2 score and EVS were both 0.86, signifying that the model explained 86% of the variance in the dependent variable, which is a strong indicator of awell-fitted model. MAPE value was 0.01, suggesting that the model was also highly accurate in relative error terms.

In contrast,theMARS modelperformedtheweakestamongallmodelsevaluated. It recorded the highest MAE (0.36), MSE (0.23), and RMSE (0.47), accompanied by a relatively low R^2 and EVS of 0.43 each. The low values of R^2 and EVS point to the model's limitedexplanatory power and suboptimal fit to the data. Although its MAPE (0.02) appears competitive, the relatively high magnitude of absolute errors undermines its predictive reliability. These results indicate that the non-linear adaptive nature of MARS did not yield substantial improvements in this context.

The ensemble-based methods, Random Forest and Gradient Boosting, demonstratedmoderate but comparable performance. Both models yielded identical MSE (0.09), RMSE (0.30),R^2(0.77),andEVS(0.77),withMAEvaluesof0.22.GBoutperformedRFintermsof MAPE (0.01 vs. 0.30), suggesting that GB provided more stable predictions relative to the magnitude of the observed values. This slight edge may be attributed to GB‘s boosting mechanism, which iteratively reduces residual error, in contrast to RF's averaging approach.

The Regression Tree model also demonstrated competitive performance, closely trailing the ensemblemodels.ItachievedalowerMAE(0.18)thanbothRFandGB,whilemaintaining the same MSE and RMSE (0.09 and 0.30, respectively). Its R^2 and EVS scores were slightly lower at 0.76, yet still indicate a strong level of explanatory power. The MAPE (0.01) was also equal to that of the OLS and GB models, implying that the Regression Tree model was quite precise in terms of percentage-based error.

From an overall perspective, the OLS model emerged as the most parsimonious and efficient model, offering the best combination of low prediction error and high explanatory power. Despite the complexity and non-linear capabilities of ensemble and tree-based models, the linear OLS model proved to be most suitable for the data at hand. This finding underscores the importance of not overlooking traditional regression methods, particularly when the underlying data structure does not exhibit significant non-linear patterns.

Table 2:
DescriptiveStatistics

Models	MAE	MSE	RMSE	MAPE	R^2-Score	EVS
OLS	0.15	0.06	0.24	0.01	0.86	0.86
MARS	0.36	0.23	0.47	0.02	0.43	0.43
RF	0.22	0.09	0.3	0.3	0.77	0.77
GB	0.22	0.09	0.3	0.01	0.77	0.77
REG	0.18	0.09	0.3	0.01	0.76	0.76

Source:Author(2024).

Table 3:
TrainedDataEvaluation Metrics

Models	OLS	MARS	Regression Tree	Random Forest	Gradient Boosting
Eva. Metrics					
MAE	0.33	0.49	0.00	0.07	0.27
MSE	0.19	0.40	0.00	0.26	0.12
RMSE	0.44	0.63	0.01	0.01	0.34
MAPE	0.02	0.03	0.00	0.84	0.02
R^2Score	0.53	0.02	1.00	0.84	0.71
EVS	0.53	0.02	1.00	0.85	0.71
Accuracy	0.76	0.51	1.00	0.85	0.80
Precision	0.72	0.50	1.00	0.84	0.79
Recall	0.82	1.00	1.00	0.84	0.80
Specificity	0.70	1.00	1.00	0.70	0.79

MCC	0.52	0.12	1.00	0.70	0.60
CohenKappa	0.51	0.03	1.00	0.70	0.60
Confusion Matrix	1052461 266 1198	56 1457 4 1460	1512 1 0 1464	1293220 228 1236	1199314 288 1176

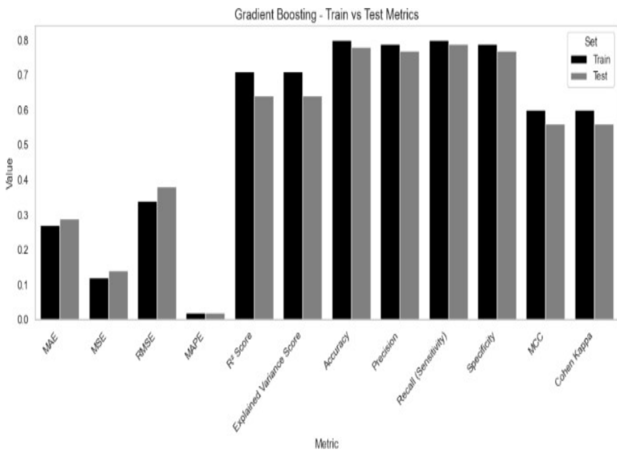
Source: Author (2024).

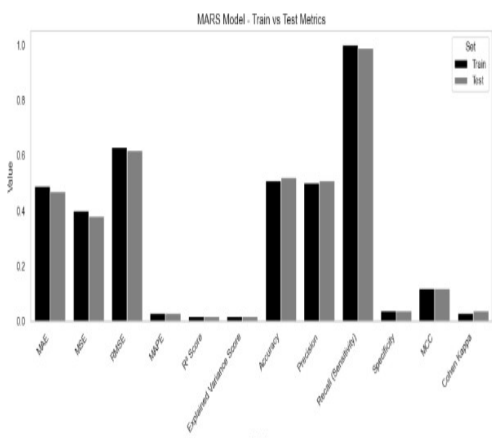
Table 4:
TestDataEvaluationMetrics:

Models	OLS	MARS	Regression Tree	Random Forest	Gradient Boosting
Eva. Metrics					
MAE	0.34	0.47	0.29	0.28	0.29
MSE	0.20	0.38	0.20	0.14	0.14
RMSE	0.44	0.62	0.45	0.37	0.38
MAPE	0.02	0.03	0.02	0.02	0.02
R^2Score	0.50	0.02	0.49	0.65	0.64
EVS	0.50	0.02	0.49	0.65	0.64
Accuracy	0.77	0.52	0.82	0.80	0.78
Precision	0.75	0.51	0.83	0.81	0.77
Recall	0.82	0.99	0.80	0.79	0.79
Specificity	0.72	0.04	0.84	0.82	0.77

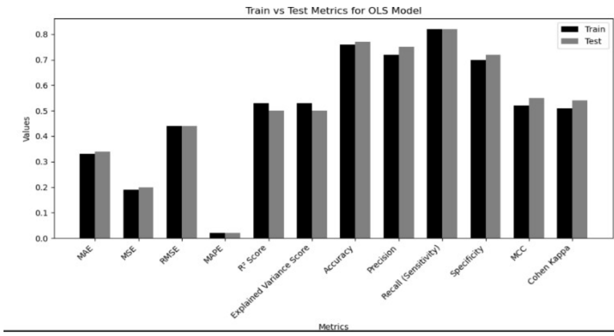
MCC	0.55	0.12	0.64	0.61	0.56
CohenKappa	0.54	0.04	0.64	0.61	0.56
Confusion Matrix	271 104 66 304	16 359 2 368	31560 74 296	30669 77 293	28986 77 293

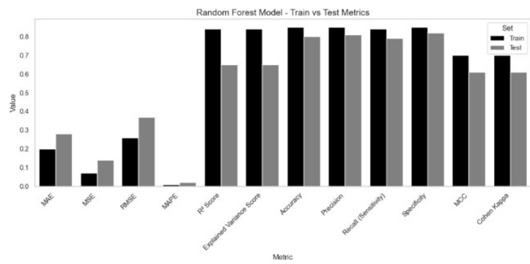
Source: Author (2024)

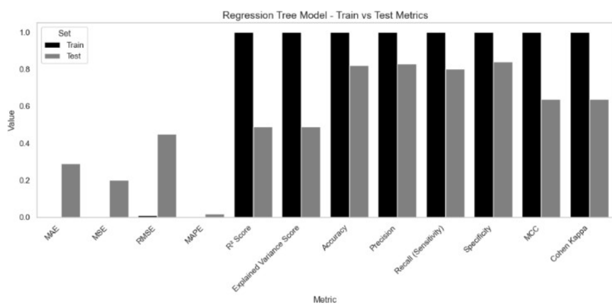


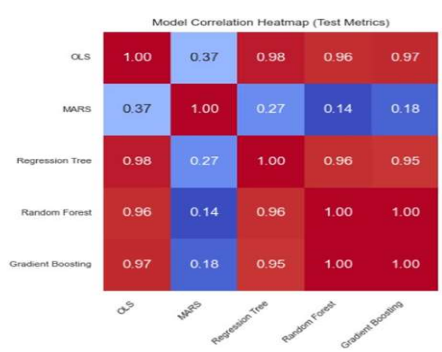


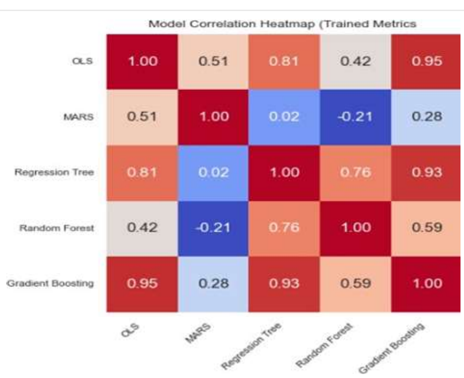
Source:Author(2024)

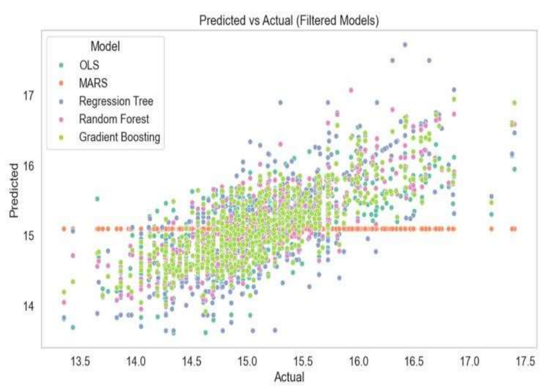

Train vs Test Metrics for OLS Model
Train
Test
Values
Metrics
MAE
MSE
RMSE
MAPE
R² Score
Explained Variance Score
Accuracy
Precision
Recall (Sensitivity)
Specificity
MCC
Cohen Kappa

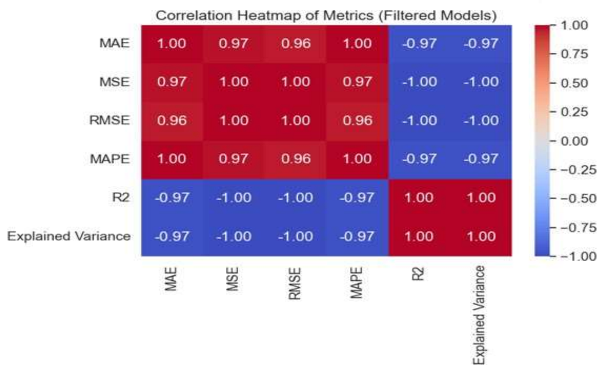

Random Forest Model - Train vs Test Metrics
Set
Train
Test
Value
Metric


Regression Tree Model - Train vs Test Metrics
Set
Train
Test
Value
Metric
MAE
MSE
RMSE
MAPE
R² Score
Explained Variance Score
Accuracy
Precision
Recall (Sensitivity)
Specificity
MCC
Cohen Kappa


Model Correlation Heatmap (Test Metrics)
OLS 1.00 0.37 0.98 0.96 0.97
MARS 0.37 1.00 0.27 0.14 0.18
Regression Tree 0.98 0.27 1.00 0.96 0.95
Random Forest 0.96 0.14 0.96 1.00 1.00
Gradient Boosting 0.97 0.18 0.95 1.00 1.00
OLS
MARS
Regression Tree
Random Forest
Gradient Boosting


Model Correlation Heatmap (Trained Metrics
OLS 1.00 0.51 0.81 0.42 0.95
MARS 0.51 1.00 0.02 -0.21 0.28
Regression Tree 0.81 0.02 1.00 0.76 0.93
Random Forest 0.42 -0.21 0.76 1.00 0.59
Gradient Boosting 0.95 0.28 0.93 0.59 1.00
OLS
MARS
Regression Tree
Random Forest
Gradient Boosting





Source:Author(2024)

PolicyImplications

The findings of this study carrysignificant implications for policymakers and stakeholders in data-driven environments. The results suggest that adopting simpler, well-established statistical models such as OLS can lead to more transparent, interpretable, and cost-effective forecasting systems, especially in policy-sensitive areas such as economic planning, environmental modeling, and financial forecasting. Furthermore, institutions engaged in predictive modeling are encouraged to build model selection frameworks based on empirical performance rather than model complexity or novelty, thus fostering efficient allocation of analytical resources.

From a policy standpoint, the findings support the integration of machine learning ensemble models, especiallyRandom Forest and Gradient Boosting, into strategic decision-makingand forecasting frameworks. These models can significantly enhance the precision of data-driven policy decisions across sectors such as finance, transportation, housing, and environmental planning. In the specific context of the automobile industry, the adoption of these models for car price prediction can aid in formulating evidence-based pricing regulations, taxation policies, and import/export tariffs. For instance, accurate car price forecasting can help policymakers design fairer vehicle taxation schemes that consider depreciation patterns and market dynamics, thereby improving equity and efficiency in policy execution.

These models' ability to quantify the influence of various predictors allows regulators to better understand how vehicle features like age, mileage, fuel type, and brand affect pricing, guidingsustainablemobilityinitiativesandconsumerprotectionpolicies.Theimproved predictive accuracy and robustness of ensemble models make them well-suited for high-stakes applications such as fraud detection in vehicle transactions, demand forecasting in urban planning, and economic modeling of transportation trends. As governments increasingly rely on intelligent systems for policy formulation and evaluation, the integration of advanced predictive analytics into public infrastructure, particularly in areas like car price monitoring - will ensure more responsive, transparent, and data-informed governance.

5.0 Conclusions

This study rigorously assessed the predictive performance of five regression models, such as the OLS, MARS, Regression Tree, Random Forest, and Gradient Boosting, for the specific task of car price prediction. Byleveraging a robust evaluation framework that integrates both error metrics and classification-based indicators (accuracy, precision, recall, specificity, Matthews Correlation Coefficient, and Cohen's Kappa), the analysis reveals distinct differences in model behavior and forecasting quality.

Among the models, ensemble-based approaches such as Random Forest and Gradient Boosting emerged as the most reliable and technically sound options, demonstrating superior generalization on unseen test data. These models efficiently capture complex, non-linear relationships within thecarpricedata,leadingto lowerprediction errors. WhileOLS remains a foundational and interpretable technique, its performance was comparatively modest, especially when the data structure exhibited higher degrees of variance or non-linearity. The MARS model, although flexible in theory, showed weak explanatory power and consistency in this context. TheRegression Tree, whileproducingperfect fit on thetrainingset, displayed symptoms of overfitting, which raises caution for its standalone deployment. In sum, for forecasting car prices, Random Forest and Gradient Boosting provide robust and accurate predictions, makingsuitable tools for both academic modelingand industryapplications such as automated valuation systems, dynamic pricing engines, and decision support platforms in automotive markets.

The implication of this result is that model performance is highly contextual and should be empirically verified rather than assumed. Simpler models like OLS not only offer interpretability but also perform competitively or even outperform complex algorithms in some scenarios. These findings advocate for a pragmatic, data-driven approach to model selection, where emphasis is placed on validation and transparency. In practical terms, stakeholders should develop and adopt model governance frameworks that integrate rigorous model testing, model interpretability, and predictive performance validation. This approach will support evidence-based decision-making, improve accountability, and enhance the reliability of forecasts across various domains.

The use of ensemble learning models, particularly Random Forest and Gradient Boosting, is highly recommended for applications requiring both precision and reliability, especially inthe domain of car price prediction. These models consistently outperformed others across a wide range of evaluation metrics on both training and testing datasets, making them particularly suitable for modeling the complex pricing structures and feature interactions inherent in automobile valuation. Their resilience to overfitting, combined with the ability to capture non-linear dynamics in vehicle characteristics such as mileage, engine type, brand, and production year, positions them as optimal choices for accurate and scalable car price forecasting systems. The OLS model, despite its relatively lower predictive performance,maystillberecommendedinuse-caseswheremodeltransparencyisparamount.Conversely, due to its weak performance and poor discriminative capacity, the MARS model shouldeither be significantly refined through advanced tuning techniques or considered unsuitable for this task. Likewise, while the Regression Tree performed well on training data, its susceptibility to overfitting warrants caution unless accompanied by pruning or robust validation protocols. For future applications in car price prediction, additional measures such as k-fold cross-validation, model regularization, and feature importance ranking should be integrated to improve the overall robustness and trustworthiness of the deployed predictive systems. The study contributes to the growing literature that calls for balanced integration of traditional and modern modeling techniques in predictive analytics. Future research could explore hybrid or ensemble configurations that combine the strengths of both linear and non- linear models, thereby enhancing predictive performance while retaining interpretability.

References

Armstrong, J. S.,&Brodie, R. (1999).Forecastingfor marketing.In G. Hooley&M. Hussey (Eds.), Quantitative methods in marketing (pp. 92–120). London: International Thomson Business Press.

Armstrong, J. S., &Brodie, R. J. (1999).Forecasting for marketing.*International Journal of Forecasting*, 15(1), 1–9.

Asur, S., &Huberman, B. A. (2010).Predicting the future with social media.*2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 1, 492–499.

Asur, S., &Huberman, B. A. (2010).Predicting the future with social media.In Proceedings of the 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology-Volume 01 (pp. 492–499).IEEE Computer Society.

Bao, H., Guo, X., Liang, J., Lan, F., Li, J., Chen, G., Mo, J., (2022). Relevance vector machine with optimal hybrid kernel function for electric vehicles ownership forecasting: the case of China. Energy Rep. 8, 988–997.

Bao, H., Zhang, Q., Wang, L., & Jiang, Y. (2022).A relational vector machine approach for predicting EV ownership using small-sample data.*Applied Energy*, 309, 118408.

Boraha, A. &Rutz, O. (2024). Enhanced sales forecasting model using textual search data: Fusing dynamics with big data. International Journal of Research in Marketing 41: 632–647

Boraha, A., &Rutz, O. (2024).Competitive online behavior and machine learning indynamic markets.*Journal of Marketing Analytics*. Advance online publication.

Du, R. Y., Kamakura, W. A., Hu, Y., &Damangir, S. (2015). Leveraging trends in online searches for product features in market response modeling. Journal of Marketing, 79(1), 29–43.

Feng, C., & Chen, W. (2021). Deep learning in transportation systems: Applications and research directions. *Transportation Research Part C: Emerging Technologies*, 129, 103196.

Feng, D., Chen, H., (2021). A small samples training framework for deep Learning-based automatic information extraction: case study of construction accident news reports analysis. Adv. Eng. Inf. 47, 101256

Han,J.,&Tang,Q.(2022).DiffusionofinnovationinsocialnetworksusingBassmodels.

InformationSystems Frontiers, 24, 1021–1035.

Han, Z.Y., & Tang, Z.J. (2022).Research on product information diffusion on social networking platform by integrating two-stage process model and improved bass model, Oper. Res. Manag. Sci. 31, 216–223.

Hong, H.-P., Cui, X.-Z., Qiao, D., (2022). Simulating nonstationary non-Gaussian vector process based on continuous wavelet transform. Mech. Syst. Signal Process. 165, 108340

Hong, X., Li, T., & Yu, J. (2022).Enhancing small-sample deep learning models via data augmentation: A wavelet-based approach.*Expert Systems with Applications*, 194, 116521.

Liu, B., Song, C., Liang, X., Lai, M., Yu Z, Ji, J. (2023). Regional differences in China's electric vehicle sales forecasting: Under supply-demand policy scenarios. Energy Policy 17, 113554

Liu, L., Liu,S.,Wu,L.,Zhu,J.,Shang,G., (2022).Forecastingthedevelopmenttrendofnew energy vehicles in China by an optimized fractional discrete grey power model. J. Clean. Prod. 372, 133708 https://doi.org/10.1016/j.jclepro.2022.133708.

Liu,M., Lu,Y., Long,S.,Bai,J.,Lian,W., (2021).Anattention-basedCNN-BiLSTMhybrid neural network enhanced with features of discrete wavelet transformation for fetal acidosis classification. Expert Syst. Appl. 186, 115714 https://doi.org/10.1016/j.eswa.2021.115714.

Liu, Z., Ma, R., & Yang, X. (2021). Hybrid CNN-BiLSTM model for predicting car ownership with wavelet-transformed input. *IEEE Transactions on Intelligent Transportation Systems*, 22(11), 6892–6903.

Øverby, H., Audestad, J.A., &Szalkowski, G.A. (2023). Compartmental market models inthe digital economy-extension of the bass model to complex economic systems, Telecommun. Policy 47, 102441.

Øverby, H., Munkvold, B. E., &Haugstveit, I. M. (2023).Modeling digital product adoption using the Bass model.*Electronic Markets*, 33, 81–94.

Qiao, W., Liu, W., Liu, E., (2021). A combination model based on wavelet transform for predicting the difference between monthly natural gas production and consumptionof. U.S.Energy. 235, 121216

Qiao, W., Sun, Y., & Zhang, K. (2021).Deep learning applications in energy and transportation systems.*Renewable and Sustainable Energy Reviews*, 135, 110128.

Zeng, Y., Zeng, Y., Choi, B., Wang, L., (2017). Multifactor-influenced energy consumption forecasting using enhanced back-propagation neural network. Energy 127, 381–396. ttps://doi.org/10.1016/j.energy.2017.03.094.

Zeng, Y., Zhang, Y., & Wang, L. (2017).Improving small-sample machine learning models through synthetic data generation.*Knowledge-Based Systems*, 128, 157–165.

Zhang, C. Tian, Y.X. &Fan, Z.P. (2022).Forecastingthe box offices of movies comingsoon using social media analysis: a method based on improved bass models, Expert Syst. Appl. 191, 116241.

Zhang, L., Wang, L., Chai, J., (2020). Influence of new energy vehicle subsidy policy on emissionreductionofatmosphericpollutants:acasestudyof Beijing,China.J.Clean. Prod. 275, 124069.

Zhang, W., Li, F., & Wang, H. (2022).Forecasting new product sales using innovation-imitation models.*Journal of Product Innovation Management*, 39(2), 202–221.

Zhu, L., Zhang, C., Zhang, C., Zhang, Z., Nie, X., Zhou, X., Liu, W., Wang, X., (2019). Forming a new small sample deep learning model to predict total organic carbon content bycombiningunsupervised learningwith semisupervised learning.Appl. Soft Comput.83, 105596.

Zhu, X., He, X., & Huang, Y. (2019). Noise-robust deep learning models for consumer prediction. *Information Sciences*, 481, 322–336.

Ma, S., Fan, Y., (2020).A deployment model of EV charging piles and its impact on EV promotion.Energy Pol. 146, 111777.

Ma, Y., Shi, T., Zhang, W., Hao, Y., Huang, J., Lin, Y., (2019). Comprehensive policy evaluation of NEV development in China, Japan, the United States, and Germany based on the AHP-EW model. J. Clean. Prod. 214, 389–402.

Rietmann,N.,Hügler,B., Lieven,T.,2020.Forecastingthetrajectoryofelectricvehiclesales and the consequences for worldwide CO2 emissions. J. Clean. Prod. 261, 121038

Sun, Y., Zhang, Y., Su, B., (2022). Impact of government subsidy on the optimal R&D and advertising investment in the cooperative supply chain of new energy vehicles.Energy Pol. 164, 112

Ding, S., Li, R., 2021. Forecasting the sales and stock of electric vehicles using a novel self-adaptive optimized grey model.Eng. Appl. Artif.Intell.100, 104148.

He, L., Pei, L., Yang, Y., (2020).An optimised grey buffer operator for forecasting the production and sales of new energy vehicles in China. Sci. Total Environ. 704, 135321.