Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Hybrid Machine Learning and Seasonal Time Series Framework for Variant-Level Monthly Car Sales Forecasting in the Automotive Industry

Tanvi kamal Rana[1] Hina Jignesh Chokshi[2]
[1]*Department of Computer Science, Parul Institute of Computer Applications, FITCS, Parul University, Vadodara, 391760, Gujarat, India*
proftanvibca@gmail.com
[2]*Department of Computer Science, Parul Institute of Computer Applications, FITCS, Parul University, Vadodara, 391760, Gujarat, India*
hina.chokshi42058@paruluniversity.ac.in

ARTICLE INFO

Received: 24 Oct 2024

Revised: 12 Nov 2024

Accepted: 26 Dec 2024

ABSTRACT

For the automotive sector to improve production planning, inventory control, and strategic decision-making, accurate automobile sales forecasting is crucial. In order to enhance variant-level monthly automobile sales prediction, this study suggests a hybrid forecasting framework that blends seasonal time series methods with machine learning models. To find feature-driven sales patterns, the study uses real dealership data and a variety of machine learning approaches, like Random Forest, XGBoost as well as Linear Regression. Simultaneously, seasonal and temporal trends are captured using the SARIMA and Facebook Prophet models. Both strategies are combined in the suggested hybrid framework to produce accurate and dependable monthly projections for specific car models. RMSE, MAE, and R2 measures are used to assess forecast accuracy. Prophet and Random Forest performed the best out of all the models. In order to provide a clear understanding of model-wise patterns, visualizations that integrate historical and projected data are created. With the help of this variant-level forecasting methodology, Original Equipment Manufacturers (OEMs) and dealerships can make better decisions about marketing and stock allocation. The suggested method helps improve forecasting accuracy and business efficiency in the automobile industry because it is flexible and scalable for practical application.

Keywords: Car Sales Forecasting, Machine Learning, Time Series Analysis, Facebook Prophet, SARIMA, Hybrid Forecasting Model

INTRODUCTION

Accurate sales forecasting is essential to operational effectiveness, strategic planning, and profitability in the cutthroat automobile sector of today. To make well-informed judgments on inventory management, production scheduling, and promotional offers, Original Equipment Manufacturers (OEMs) and auto dealerships must comprehend and predict monthly vehicle demand at the variant level.

Moving averages and ARIMA are two statistical and time series models that have historically been used extensively in forecasting. These models are good at capturing broad trends, but they frequently perform poorly when data shows intricate, nonlinear patterns or involves several influencing factors, such as buyer type, discount rates, fuel type, and geographic demand. Furthermore, festival season surges, variant-specific purchasing patterns, and changing consumer preferences—all of which are critical in the Indian automobile market—are not automatically taken into account by these models.

Machine learning (ML) techniques have become strong substitutes to overcome these drawbacks. By using a variety of characteristics, including booking amount, discount, fuel description, and model variant, models such as Random Forest, XGBoost, and Linear Regression can learn from past datasets and uncover hidden trends. The seasonal trends and time-based swings that are essential for predicting car sales—especially in areas where monthly and festival-driven patterns are prevalent—are frequently overlooked by these models, which regard time as a static characteristic.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Conversely, seasonal time series models, like Facebook Prophet and SARIMA, are made expressly to identify and simulate seasonality, trends, and time-based dependencies. Prophet, created by Meta (previously Facebook), works well for recording irregular trends, holidays, and monthly patterns. However, time series models usually only use univariate data and do not account for multidimensional feature interactions, which limits their capacity to forecast sales that are influenced by promotional strategies or consumer behavior.

On the other hand, periodic time series models, like SARIMA as well as Facebook Prophet, are specifically designed to detect and model trends, seasonality, and time-based relationships. Prophet, developed by Meta (formerly Facebook), is useful for documenting monthly patterns, holidays, and erratic tendencies. However, time series models are limited in their ability to predict sales that are impacted by customer behavior or promotional techniques since they often only employ univariate data and fail to take multidimensional feature interactions into account.

RMSE, MAE, and R2 Score are among the measures used to assess the models in order to guarantee forecasting accuracy and business applicability. The study's findings not only increase the precision of auto sales forecasts but also serve as a decision-support tool for dealerships and OEMs, assisting them in effectively managing demand variations, planning stock distribution, and optimizing promotional tactics.

This research saturates a substantial vacuum in the present literature by providing a forecasting system that is application-focused, scalable, and explicable, all of which are in line with actual industry demands. The forecasting solution it designs is not only technically accurate but also business-aware and variant-specific, making it extremely relevant to the car retail ecosystem. This makes it a methodological and practical contribution.

RELATED WORK

Sales forecasting is vital for the automotive industry, enabling manufacturers and dealerships to optimize production, inventory, and financial planning. Traditional methods like time-series analysis and regression often fail to capture the complex relationships between factors influencing sales, such as customer preferences, financing options, fuel type, and car model. Machine learning offers a more accurate predictive approach by analyzing historical data and identifying hidden patterns.

The study by O. Access, A. C. Das, and S. Akter (2024) focuses on optimizing retail demand forecasting by assessing machine learning models together with LSTM and Gradient Boosting. Using historical sales data, the research employs data preprocessing, feature selection, and hyperparameter tuning. Models are assessed based on MAE, RMSE, and R^2 scores. (Access et al., 2024)

In (Anggriawan & Gunawan, 2022) explore bicycle sales prediction with the K-Means algorithm for data mining. The work objective to segment sales data to identify purchasing patterns and optimize inventory management. The methodology includes data preprocessing, clustering, and pattern analysis. Implementing K-Means helps group sales trends effectively. The research utilizes Python and statistical tools, evaluating cluster validity through SSE and silhouette scores. Strengths include practical applicability and robust clustering techniques.

Aguilar-Palacios et al. (2020) use differentiating clarifications and gradient boosting in order to enhance cold-start promotional sales forecasting. The study uses SHAP for interpretability, feature selection, and model development. RMSE and MAE are utilized for evaluation, and Python and Scikit-learn are used. Clarity and practical applicability are among its advantages..(Aguilar-Palacios et al., 2020)

In (Hülsmann et al., 2012) propose a general sales forecast model for automobile markets, aiming to analyze market trends and improve demand predictions. The methodology integrates statistical modeling and machine learning techniques to assess sales patterns. Implementation includes data preprocessing, trend analysis, and forecasting model development. Strengths include adaptability to different markets and robust analysis.

In (Rožanec et al., 2021) investigate automotive OEM demand forecasting by comparing various forecasting algorithms and strategies. The study aims to improve demand prediction accuracy for original equipment manufacturers. The methodology includes data preprocessing, model selection, and performance evaluation. Implementation involves testing multiple algorithms, including time-series models and machine learning

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

techniques. Python and statistical tools are used, with RMSE and MAPE as evaluation metrics. Strengths include extensive model comparison and practical applicability.

In (Panarese et al., 2022) develop and test a machine learning-based sales forecasting platform using Gradient Boosting. The goal of the study is to improve prediction accuracy and support data-driven decision-making. The strategy encompasses data preparation, attribute selection, and model training. Implementation focuses on Gradient Boosting for trend identification and demand estimation. Python, Scikit-learn, and XGBoost are utilized, with RMSE and MAE as evaluation metrics.

Madhuvanthi et al. (2019) use machine learning to improve their forecasts of automobile sales. Key sales factors, gathering data, and model training using decision trees and regression models are the primary subjects of the study. RMSE and MAE are used to evaluate Scikit-learn and Python. Model comparisons and practical applicability are among its advantages.(Madhuvanthi et al., 2019)

Subramanian et al. (2020) propose a combined model for forecasting Indian automobile sales, aiming to improve prediction accuracy using multiple data sources. The methodology involves data gathering, transformation and model training. Implementation integrates statistical and machine learning approaches to capture market trends. Tools like Python and econometric models are used, with RMSE and MAPE for evaluation.(Subramanian et al., 2020)

In (Shymanskyi & Liaskovets, 2023) propose a cascade model for predicting car sales price and time, aiming to enhance forecasting accuracy. The approach comprises multi-stage data preprocessing, feature engineering, and model training. Implementation integrates machine learning techniques to sequentially predict price and sales duration. Python, Scikit-learn, and ensemble models are used, with RMSE and MAPE for evaluation. [9]

Cheng and Wang (2023) focus on car sales prediction, aiming to enhance forecasting accuracy consuming machine learning procedures. The approach comprises data preprocessing, feature selection, and model training. Implementation involves applying regression-based and ensemble learning models to analyze sales trends. Python, Scikit-learn, and statistical analysis tools are used, with RMSE and MAPE as evaluation metrics.(Cheng & Wang, 2023)

In [22] focus on sales forecasting in the automobile industry, aiming to improve demand prediction accuracy. The methodology involves data collection, preprocessing, and statistical forecasting techniques. Implementation utilizes regression models and time-series analysis to predict sales trends. Tools like MATLAB and Excel are used, with evaluation metrics including MAPE and RMSE. Strengths include a structured approach to forecasting with real-world applicability.

Gulsun and Aydin (2024) optimize the Extreme Gradient Boosting (XGBoost) algorithm using metaheuristic techniques to enhance sales forecasting accuracy. The study aims to improve prediction performance by fine-tuning hyperparameters through metaheuristic algorithms. The procedure comprises data preprocessing, feature selection, and model optimization. Implementation integrates XGBoost with optimization techniques like Genetic Algorithms. Python and Scikit-learn are used, with RMSE and MAE as evaluation metrics. Strengths include improved model efficiency.[23]

Hanumanthappa and Sarakutty T K (2011) apply data mining to predict car manufacturing trends, aiding decision-making and production planning. Using classification and clustering techniques, the study employs WEKA and SQL, evaluating performance through accuracy and precision.[24]

In [25] propose a machine learning approach for car popularity prediction, aiming to analyze factors influencing consumer preferences. The approach contains data preprocessing, feature selection, and model preparation. Implementation includes classification algorithms for example Decision Trees and Random Forest to expect car popularity based on specifications and market trends. Python and Scikit-learn are used, with accuracy and F1-score as evaluation metrics. Strengths include effective feature analysis.

Saraswathi et al. (2021) focus on sales prediction using machine learning approaches to improve forecasting precision. The work goals to identify key sales factors and improve predictive performance. The procedure encompasses data preprocessing, feature assortment, and model training. Enactment comprises machine learning

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

algorithms like Linear Regression and Random Forest. Python and Scikit-learn are used, with RMSE and MAPE as evaluation metrics. Strengths include a comparative model analysis. [26]

OBJECTIVES

The intention of this analysis aims to develop a hybrid forecasting model that improves the automotive industry's ability to predict monthly automobile sales at the variation level. By including characteristics like invoice amount, model, fuel type, booking amount, buyer type, discount, and invoice date, this study seeks to use actual dealership sales data. The study models intricate, feature-driven patterns affecting auto sales by implementing and assessing conventional machine learning techniques including Random Forest, XGBoost, and Linear Regression. Seasonal time series models, like Facebook Prophet and SARIMA, are used in parallel to capture monthly trends, seasonality, and holiday influences that traditional machine learning algorithms are unable to adequately model.

One of the main goals is to create a hybrid forecasting framework that combines the two methods, utilizing time series models' capacity to identify temporal correlations and machine learning's prowess in managing multidimensional data. The framework provides a clear and visual view of past and future sales trends by generating projections on a month-by-month basis, based on variants. Metrics like RMSE, MAE, and R2 Score are used to assess the predicting performance. The ultimate goal of the project is to give OEMs and dealerships a scalable decision-support system that allows for data-driven planning of production, inventory, and marketing strategies.

METHODOLOGY

In order to forecast monthly car sales at the variation level, this study uses a hybrid forecasting methodology that combines seasonal time series models with machine learning.

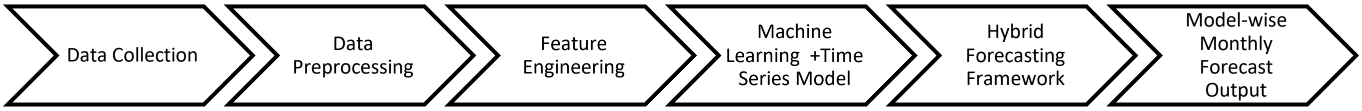

Figure 1: Hybrid Forecasting Framework – ML + Time Series Integration

Data preprocessing and feature engineering, which includes the development of new variables like Discount Rate, Vehicle Age, and Sale Month, are the first steps in the process. To find intricate patterns, machine learning models lsuch as Random Forest, XGBoost, as well as Linear Regression are trained on feature-based data. In parallel, monthly sales aggregates are subjected to time series models such as Facebook Prophet and SARIMA in order to identify patterns and seasonality.

The advantages of both strategies are then combined to create a hybrid framework. Dealerships are able to make data-driven decisions about manufacturing, promotions, and inventories thanks to the model-wise monthly projections generated by this platform. RMSE, MAE, and R2 are used to assess forecast performance, and comparisons between past and anticipated trends are shown through visualizations.

The proposed model employs statistical and mathematical formulations to optimize predictive accuracy:

1. Feature Selection & Correlation Analysis

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} \qquad (1)$$

Where r is the Pearson correlation coefficient, X and Y are variables influencing sales.

2. Regression-Based Sales Prediction

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \quad (2)$$

Where Y is the predicted sales, Xi are independent variables, βi are coefficients, and ϵ is the error term.

3. Machine Learning Classification for Sales Categories

   Logistic Regression

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\sum \beta_i X_i)}} \quad (3)$$

Support Vector Machine (SVM) Decision Boundry

$$f(x) = w^T X + b \quad (4)$$

Where www is the weight vector, X is the input feature vector, and b is the bias term.

4. Error Metrics for Model Appraisal

Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \hat{Y}_i| \quad (5)$$

Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \quad (6)$$

F1 Score for Classification

$$F1 = 2 \times \frac{Precision \ \times Recall}{Precision \ \times Recall} \quad (7)$$

The integration of these statistical and mathematical models ensures an accurate, data-driven approach to forecast automobile sales, enabling optimized decision-making in inventory, marketing, and financial planning.

RESULT AND DISCUSSION

Using actual automobile sales data, the efficiency of the recommended hybrid expecting framework was measured with an emphasis on monthly, variant-level forecasts. Both time series forecasting methods like SARIMA and Facebook Prophet in addition to machine learning models similar to Linear Regression, Random Forest, and XGBoost were used in the investigation. With its greatest R2 score (0.9967) and lowest error rates, Random Forest had the best performance in predicting automobile sales, as seen by the data in Table 1. Due to its incapacity to identify nonlinear patterns, Linear Regression produced subpar results, although XGBoost also demonstrated strong performance. Facebook Prophet performed better than SARIMA among time series models because it was able to better capture seasonal trends. All things considered, the hybrid strategy that used time series models and machine learning worked best for projecting monthly auto sales at the variation level.

Table 1: Forecasting Model Performance Comparison

Model	RMSE	MAE	R^2 Score	Key Observation
Linear Regression	2,01,145.13	1,36,541.80	0.4473	Poor performance, unable to capture complexity
Random Forest	**15,513.65**	**6,211.54**	**0.9967**	Best performance among all models
XGBoost	21,127.55	10,927.58	0.9939	Robust and accurate
SARIMA	~23,000+	~9,800+	~0.9891	Good for trend but less responsive to features
Facebook Prophet	~18,000+	~8,500+	~0.9925	Captures seasonality and trend effectively

The findings demonstrate that the accuracy of predicting auto sales is much increased when machine learning and time series models are combined. Among ML models, Random Forest produced the best accurate predictions, and

Facebook Prophet well represented seasonality. The hybrid technique made it possible to produce accurate monthly projections at the variation level, providing OEMs and dealerships with useful information to better plan their marketing and inventory plans.

REFERENCES

[1] Access, O., Das, A. C., & Akter, S. (2024). OPTIMIZING RETAIL DEMAND FORECASTING : A PERFORMANCE EVALUATION OF MACHINE LEARNING MODELS INCLUDING LSTM AND GRADIENT. 06, 67–80.

[2] Aguilar-Palacios, C., Munoz-Romero, S., & Rojo-Alvarez, J. L. (2020). Cold-Start Promotional Sales Forecasting through Gradient Boosted-Based Contrastive Explanations. IEEE Access, 8, 137574–137586. https://doi.org/10.1109/ACCESS.2020.3012032

[3] Anggriawan, I., & Gunawan, W. (2022). Implementation of Data Mining Using K-Means Algorithm for Bicycle Sales Prediction. ILKOM Jurnal Ilmiah, 14(3), 284–293. https://doi.org/10.33096/ilkom.v14i3.1291.284-293

[4] Cheng, M., & Wang, Z. (2023). Car Sales Prediction. 06, 3271–3273.

[5] Hülsmann, M., Friedrich, C. M., Reith, D., & Borscheid, D. (2012). General Sales Forecast Model for Automobile Markets and their Analysis General Sales Forecast Models for Automobile Markets and their Analysis. Transactions on Machine Learning and Data Mining, 5(2), 65–86. https://www.researchgate.net/publication/230626894

[6] Madhuvanthi, K., Nallakaruppan, M. K., Senthilkumar, N. C., & Siva Rama Krishnan, S. (2019). Car sales prediction using machine learning algorithmns. International Journal of Innovative Technology and Exploring Engineering, 8(5), 1039–1050.

[7] Panarese, A., Settanni, G., Vitti, V., & Galiano, A. (2022). Developing and Preliminary Testing of a Machine Learning-Based Platform for Sales Forecasting Using a Gradient Boosting Approach. Applied Sciences (Switzerland), 12(21). https://doi.org/10.3390/app122111054

[8] Rožanec, J. M., Kaži˘, B., Škrjanc, M., & Fortuna, B. (2021). Automotive-OEM-demand-forecasting-A-comparative-study-of-forecasting-algorithms-and-strategiesApplied-Sciences-Switzerland (1).pdf.

[9] Shymanskyi, V., & Liaskovets, V. (2023). Cascade Model for Price and Time of Car Sales Prediction. CEUR Workshop Proceedings, 3641, 152–167.

[10] Subramanian, K., Othman, M., Sokkalingam, R., Thangarasu, G., & Kayalvizhi, S. (2020). An Integrated Model for Forecasting Indian Automobile. International Journal on Advanced Science, Engineering and Information Technology, 10(6), 2593–2598. https://doi.org/10.18517/ijaseit.10.6.8475

[11] Access, O., Das, A. C., & Akter, S. (2024). OPTIMIZING RETAIL DEMAND FORECASTING : A PERFORMANCE EVALUATION OF MACHINE LEARNING MODELS INCLUDING LSTM AND GRADIENT. 06, 67–80.

[12] Aguilar-Palacios, C., Munoz-Romero, S., & Rojo-Alvarez, J. L. (2020). Cold-Start Promotional Sales Forecasting through Gradient Boosted-Based Contrastive Explanations. IEEE Access, 8, 137574–137586. https://doi.org/10.1109/ACCESS.2020.3012032

[13] Anggriawan, I., & Gunawan, W. (2022). Implementation of Data Mining Using K-Means Algorithm for Bicycle Sales Prediction. ILKOM Jurnal Ilmiah, 14(3), 284–293. https://doi.org/10.33096/ilkom.v14i3.1291.284-293

[14] Cheng, M., & Wang, Z. (2023). Car Sales Prediction. 06, 3271–3273.

[15] Hülsmann, M., Friedrich, C. M., Reith, D., & Borscheid, D. (2012). General Sales Forecast Model for Automobile Markets and their Analysis General Sales Forecast Models for Automobile Markets and their Analysis. Transactions on Machine Learning and Data Mining, 5(2), 65–86. https://www.researchgate.net/publication/230626894

[16] Madhuvanthi, K., Nallakaruppan, M. K., Senthilkumar, N. C., & Siva Rama Krishnan, S. (2019). Car sales prediction using machine learning algorithmns. International Journal of Innovative Technology and Exploring Engineering, 8(5), 1039–1050.

[17] Panarese, A., Settanni, G., Vitti, V., & Galiano, A. (2022). Developing and Preliminary Testing of a Machine Learning-Based Platform for Sales Forecasting Using a Gradient Boosting Approach. Applied Sciences (Switzerland), 12(21). https://doi.org/10.3390/app122111054

[18] Rožanec, J. M., Kaži˘, B., Škrjanc, M., & Fortuna, B. (2021). Automotive-OEM-demand-forecasting-A-comparative-study-of-forecasting-algorithms-and-strategiesApplied-Sciences-Switzerland (1).pdf.

[19] Shymanskyi, V., & Liaskovets, V. (2023). Cascade Model for Price and Time of Car Sales Prediction. CEUR Workshop Proceedings, 3641, 152–167.
[20] Subramanian, K., Othman, M., Sokkalingam, R., Thangarasu, G., & Kayalvizhi, S. (2020). An Integrated Model for Forecasting Indian Automobile. International Journal on Advanced Science, Engineering and Information Technology, 10(6), 2593–2598. https://doi.org/10.18517/ijaseit.10.6.8475