Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Unlocking business insights with big data analytics and predictive AI: Discovering hidden patterns for accurate sales forecasting

Taher M. Ghazal [1], Nabil El Kadhi [2], Munir Ahmad [3, 4, *]

[1]College of Arts and Science, Applied Science University, P.O. Box 5055, Manama, Bahrain
[2]VPAA and Computer Science Department, Applied Science University, P.O. Box 5055, Manama, Bahrain
[3]School of Computer Science, National College of Business Administration and Economics, Lahore 54000, Pakistan
[4]University College, Korea University, Seoul 02841, South Korea

ARTICLE INFO

Article history:
Received 17 November 2024
Received in revised form
19 April 2025
Accepted 1 August 2025

Keywords:
Sales forecasting
Big data
Predictive AI
Machine learning
Business analytics

ABSTRACT

In today's digital era, businesses are increasingly adopting innovative approaches to gather valuable data for informed decision-making and maintaining competitiveness. This study examines the application of big data analytics and predictive artificial intelligence (AI) in sales forecasting, a task that remains challenging but essential for effective demand planning and resource allocation. Traditional forecasting methods often fall short in dynamic market environments, whereas advanced techniques offer greater accuracy. Using real-world data, this research employs machine learning algorithms to uncover hidden patterns and generate reliable sales predictions. A predictive model based on the XGBoost algorithm was developed and achieved a high R^2 score of 0.94, with cross-validation yielding a consistent mean score of 0.94 (SD = 0.02), indicating strong predictive power and stability. The findings demonstrate the effectiveness of big data and predictive AI in improving forecast accuracy and supporting data-driven business decisions. This study highlights the practical value of integrating advanced analytics into sales forecasting processes for enhanced strategic planning.

1. Introduction

In recent years, the term big data has gained immense popularity, specifically in the business world. The rapid growth of digital information has led to a vast amount of data being generated every day. With the help of modern technologies, companies can now collect, store, and process these data sets to gain valuable insights that were previously impossible. Businesses have started to realize the potential benefits of leveraging big data analytics and predictive AI in their operations, and the area of sales forecasting is no exception (Rashidi et al., 2025; Gupta et al., 2019).

Sales forecasting is a critical task for businesses, as it enables them to anticipate future demand and allocate resources effectively. Conventional methods of sales forecasting have been used in the past, but these methods often fail to provide sufficient accuracy, particularly in an abruptly changing environment. In contrast, big data analytics can help uncover hidden patterns and make accurate predictions by business organizations (Ahaggach et al., 2024).

The utilization of big data analytics in sales forecasting is one of the most attempted areas of research. These technologies offer businesses the opportunity to leverage vast amounts of data to gain valuable insights and make informed decisions. The use of machine learning algorithms in sales forecasting is particularly promising, as these algorithms can identify complex patterns in data that are difficult for humans to discern (Pavlyshenko, 2019).

The prime objective of this research study is to explore the application of big data analytics and predictive AI in sales forecasting. Particularly, we will examine how these technologies can be utilized to improve accuracy. We will provide a case study that uses machine learning (ML) algorithms with real-world data to demonstrate how predictive AI is applied in sales forecasting.

Our study will explore the application of XGBoost. Considering its excellent precision, scalability, and

* Corresponding Author.
Email Address: munirahmad@korea.ac.kr (M. Ahmad)
https://doi.org/10.21833/ijaas.2025.08.022
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-5240-0984

reliability, XGBoost has effectively been used in many industries. This study is exclusively on the performance evaluation of XGBoost, particularly for the accurate prediction of sales volume.

The PakWheels' Listing Dataset (Kaggle, 2024) is one of Pakistan's prominent auto-listing datasets. It comprises a variety of information regarding vehicle specifications, costs, and sales history. In the data preprocessing phase, we will remove data errors and inconsistencies. Then, we will transform the data into a suitable format for analysis. Lastly, key features will be selected to focus on important variables. These processes are essential for guaranteeing high-quality datasets.

To gain an in-depth understanding of key patterns and correlations among the variables, we will perform explanatory data analysis (EDA). Next, we will use XGBoost regression to predict sales based on historical data. The model will be trained and tested using cross-validation techniques. The results would help not only to ensure the effectiveness of XGBoost in the prediction of sales but also to provide valuable insights into the significant features that contribute to accurate predictions.

In conclusion, this paper aims to explore the potential benefits of big data analytics along with predictive AI in sales forecasting. The case study using the PakWheels' Listing Dataset (Kaggle, 2024) will demonstrate the use of XGBoost in sales forecasting and provide valuable insights for businesses to make informed decisions. This study contributes to the growing body of research on the application of big data analytics and predictive AI in various industries and emphasizes the importance of leveraging these technologies to gain a competitive advantage in the market.

2. Literature review

The use of machine learning algorithms in sales forecasting has achieved remarkable attention in recent years. Big data analytics and predictive AI support business managers in making smart decisions. Various researchers have conducted studies on sales forecasting models using ML techniques such as Random Forest, Gradient Boosting, XGBoost, etc.

These models utilize historical sales data and other key factors, such as weather, promotions, and holidays, to forecast future sales. The accuracy improvement capabilities of Machine learning techniques have induced researchers to explore their application in various industries. This literature review will present existing studies on sales forecasting using ML models, including their implementation in various business sectors.

Yun et al. (2022) provided a comprehensive survey of the current trends and applications of predictive analytics. The authors emphasize the importance of predictive analytics, helpful in making smart decisions by supporting data from various sources. The authors elaborate on the various types of predictive analytics techniques, such as machine learning, deep learning, and other statistical models. They also describe their applications in various areas, such as transportation, energy, and healthcare. Further, identify the challenges relating to predictive analytics. These include data quality and privacy concerns, and provide potential solutions to overcome these challenges. The paper provides an in-depth understanding of predictive analytics in smart city planning.

Iyengar et al. (2023) recommended a novel approach for forecasting cardiovascular disease evaluation. Utilizing the Framingham dataset, the researchers applied machine learning to predict cardiovascular disease risk. They explain the methodology used to develop the predictive model and the performance evaluation measures used. According to the findings of this study, the proposed model has an accuracy of 90.87%. This paper gives useful information about the application of big data analytics in the healthcare industry. Potential advantages of Machine learning risk assessments and predictions have also been highlighted.

Forecasting accurate demand is the key requirement for the management of the supply chain. Big data analytics substantially improve decision-making capabilities. It helps businesses to more efficiently handle inventory, production schedules, and logistics for transportation.

Seyedan and Mafakheri (2020) discussed in detail the big data prediction processes helpful in accurate projection of accurate forecasting. They present the application of ML and big data in supply chain management in a summarized manner. The authors highlight the importance of AI tools for researchers and business professionals. Several advantages and disadvantages of ML techniques for demand forecasting have been discussed by the authors. A variety of data sources, including social media, historical sales, and meteorological data covered in this study.

The authors indicate multiple research gaps in the area. They underscore the importance of combining different sources of data to develop a comprehensive perspective of demand patterns. They suggest hybrid models, which combine multiple forecasting methods to get higher accuracy. The study also highlights the importance of incorporating external factors, such as economic indicators and geopolitical events, into forecasting models for better decision-making.

Punia and Shankar (2022) proposed a decision support system for demand prediction based on deep learning technology. They describe the limitations of standard projection techniques, such as the inability to manage extremely complicated data adequately. To address these challenges, the authors provide a novel deep learning model that combines predictive analytics with more advanced techniques. The model's performance was tested, and it outperformed all the other techniques used for demand prediction. The researchers present comparative analyses of traditional versus

concurrent methodologies of forecasting. The authors also provide a comprehensive overview of the advantages and limitations. The study emphasizes the deep learning potential of demand forecasting.

Cadavid et al. (2018) presented an assessment of prominent trends in ML for demand and sales volume prediction. They highlight the importance of accurate prediction in supply chain management. The authors present various machine learning techniques and explain how these techniques can be utilized to improve demand and sales accuracy.

The study also explores key issues involved in ML-based prediction. These include dataset quality and computational complexity. The authors provided detailed information helpful for increasing the prediction accuracy rate. The study also provides a summary of current trends in AI applications.

Smith and Côté (2022) proposed predictive analytics for accurate sales prediction in pop-up shop setups. This research study focuses on a pop-up shop offering handmade jewelry. The authors point out the difficulties faced by small merchants. The major problems they discussed included the limited availability of data and anticipating consumer behavior.

The authors review the process of data preparation and feature engineering in detail. They forecast revenues using regression and neural networks. The results reveal that the predictive model performs much better than the standard sales forecasting methods.

The report highlights the advantages of predictive analytics, such as lower inventory costs and increased customer satisfaction. However, the authors recognize constraints, such as the requirement for huge datasets and the possibility of bias in machine learning algorithms.

Caglayan et al. (2020) explored retail store sales forecasting via machine learning techniques. They explain the role of accurate forecasting for efficient inventory management and its impact on customer satisfaction.

The dataset comprises two years of sales data from various stores and developed ANN models. The Artificial neural network is a widely used technique for managing complex relationships among the seasonality, promotions, and product attributes of sales data.

The results reveal that ANN-based models are most suitable for sales forecasting. The authors propose using an ANN to recognize nonlinear sales patterns for improvement in forecasting accuracy. The authors also suggest that future research could explore combining ANN with other machine learning techniques for even better results.

All this research and case studies collectively highlight the effectiveness of advanced analytics for supply chain management. Big data, machine learning, and deep learning are substantially improving demand and sales forecasting. By addressing challenges like data integration and algorithm optimization, researchers are paving the way for continued advancements in this field. Each study offers valuable insights and lays the groundwork for future innovations in forecasting techniques.

Faheem et al. (2024) presented an overview of Artificial intelligence's impact on business practices. They describe how AI has entirely changed financial forecasting processes. They highlighted how AI enables better risk evaluation and decision-making. AI models learn by modifying themselves over time, making them more in line with the changing market conditions. However, the study also discussed the challenges of AI-enabled financial prediction, such as data security problems, model interpretability, and ethical issues associated with automated decision-making. AI in predictive analytics is helpful for financial analysts and institutions to support decision-making and obtain better results in ever-changing markets.

Ayub et al. (2020) highlighted the importance of big data Analytics for accurate forecasting. They present a model that uses deep learning with supervised machine learning approaches to forecast power loads. The proposed model involved three steps: feature selection, extraction, and classification.

A combination of Random Forest (RF) and Extreme Gradient Boosting (XGB) is utilized to determine feature relevance. Hybrid approaches prioritize the most important and relevant elements during feature selection. Load forecasting is performed using Support Vector Machines (SVM) and a combination of GRU and CNN. Our enhanced techniques, CNN-GRU-EWO and SVM-GWO, achieve accuracy rates of 96.33% and 90.67%, respectively.

Okeleke et al. (2024) conducted a study on consumer behavior. They explore AI models using big data to evaluate historical data to identify patterns and thereby make accurate predictions about future trends. This capability is especially relevant in the rapidly changing market environment. This study focuses on a variety of AI methods, including deep learning, machine learning, and natural language processing models. The authors emphasize their ability to improve prediction accuracy. For example, complicated and large-scale data can be processed by an ML model to estimate customer demand. They also discussed the Challenges generally experienced in the Implementation of AI-driven predictive Analytics. Data quality and effective integration are the most critical challenges.

Fathima et al. (2024) highlighted the advantages and applications of big data analytics of AI-based demand forecasting. AI has transformed demand forecasting into ERP systems. Big data analytics help improve accuracy in predicting future patterns. The authors examine the influence of AI-based predictive analytics on demand forecasting in ERP systems. They focused on many areas such as fashion retail, energy management, biopharmaceuticals, and transportation. AI-driven demand forecasting offers benefits such as predicting client requirements, optimizing inventory levels, and making data-driven

decisions, giving businesses a competitive advantage in the market. Their study highlights the significance of incorporating AI into ERP systems for firms looking to improve decision-making.

3. Methodology

Our prime objective in conducting this study is to explore the effectiveness of data analytics and applications for the accurate prediction of sales for any business organization.

3.1. Dataset

Relevant and diversified datasets perform a key role in the development of the ML model. In our proposed model, we prefer to use a dataset from Pakistan's largest automotive portal, PakWheels, which was sourced from Kaggle (2024). It contains ample attributes of vehicles. In the pre-processing phase, we removed some features that were deemed redundant to our study. The reason is to enhance prediction accuracy as well as to reduce the dataset's dimensionality. Furthermore, we handled missing data by either eliminating instances with missing values or imputing the missing values with an appropriate value based on the feature distribution and proportion of missing values.

3.2. Feature engineering

Under the feature engineering phase, we apply data augmentation techniques by combining and modifying existing features to create new features. It helps to improve the performance of the prediction model.

Next to feature engineering, we conducted EDA to gain an in-depth understanding of the dataset, including features and patterns. We performed EDA by conducting bivariate analysis and univariate analysis. The purpose of bivariate analysis is to evaluate the connection between two variables. On the other hand, in univariate analysis, we examine the distribution of each variable. These studies helped us find anomalies, trends, and patterns in the dataset and gave us awareness about the variables that were crucial to our investigation.

3.3. EDA

It is frequently used by data analysts to analyze the characteristics of datasets. EDA is also helpful to visualize the structure and understand the relationships between variables. We perform this analysis to identify the relationships between different variables. It helps us to understand the existing trends and patterns within the data. We conducted EDA with three different kinds of analysis, which include Univariate, Bivariate, and Multivariate analysis. Under Univariate analysis, we obtained information about the distribution of individual variables. In Bivariate analysis, we identified the relationships between two variables. These analyses helped us to recognize key variables that have a significant impact on the target variable. This was an important step in developing a predictive model. EDA also helps to identify and manage the missing values. With the help of EDA, we can identify which variables had a strong relationship with the target variable.

EDA has been extensively used to understand the dataset. There are three key visualizations. Fig. 1 exhibits the first visualization in the form of a bar chart. It displays the top ten car manufacturers in the dataset, along with the number of vehicles listed for each manufacturer. The X-axis represents the manufacturer names, while the Y-axis shows the number of vehicles listed. The objective of this visualization is to understand the popularity of different car manufacturers and their relative market share.

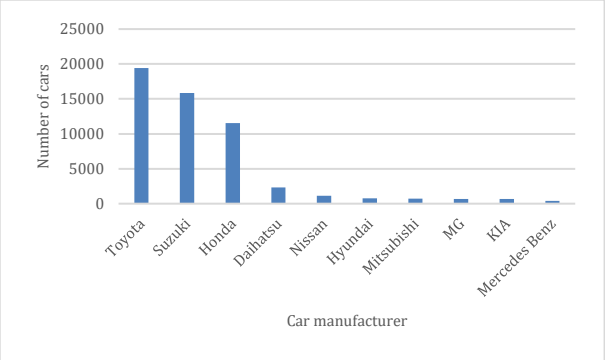

Fig. 1: Top ten car manufacturers

Next, we observe the second visualization, as shown in Fig. 2; it is a pie chart that displays the top ten car models in the dataset. The chart shows the proportion of each model in the dataset (Corolla being the top model). This visualization helped us to understand the popularity of different car models, which indicates market demand.

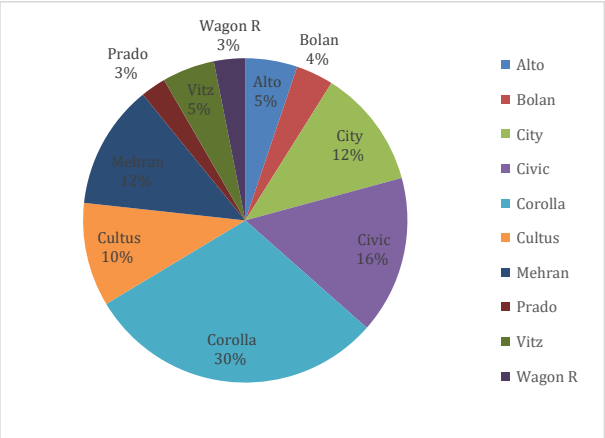

Fig. 2: Top car models in the data

The third visualization is exhibited in Fig. 3, which presents a bar chart illustrating the number of vehicle sales across different cities in Pakistan. This chart provides a clear comparison of sales volumes by region, highlighting variations among provinces

and cities. By visualizing the data in this manner, it becomes easier to identify which areas show higher or lower sales activity.

Data visualization is a key step in understanding the relationships among variables, as it reveals trends and patterns within the dataset. Such insights serve as the foundation for deeper analysis and modeling.

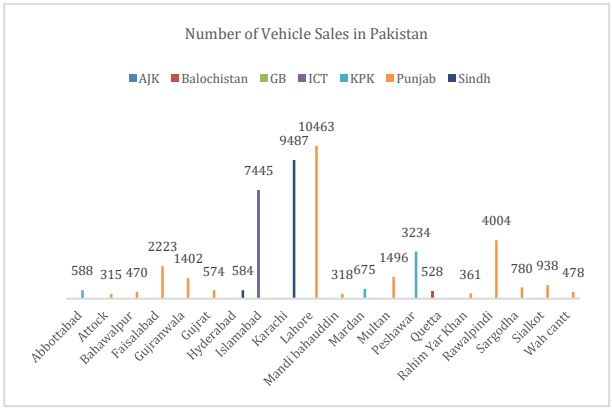

Fig. 3: Number of vehicle sales across Pakistan

3.4. Univariate analysis

It is a statistical analysis frequently conducted by data scientists to gain detailed information about a single variable independently. The information includes the tendency, spread, and shape of the data distribution. The objective is to understand the data. Univariate analysis provides useful information about each variable separately, irrespective of other variables. This technique is very helpful to identify patterns and relationships within the data.

Univariate analysis is most suitable to determine whether the data is symmetrically distributed or not. Whether it is multimodal or there are any outliers exist. This information is necessary for the appropriate handling of the variable in further analysis. It is also useful to identify any issues with the data, such as missing or extreme values.

Fig. 4 shows a pie chart of the top ten cities showing the number of cars sold.

The pie chart shows the proportion of car sales in each city. The city with the highest car sales can be observed from the chart. It also presents a comparison with other cities. This visual is a classic example of univariate analysis. It focuses on the distribution of a single variable. In this example, the variable is the number of car sales in cars in various cities. The objective of this analysis is to gain an understanding of car sales distribution across different cities. Visual also identifies the city with the highest car sales and compares the car sales across different cities.

Our findings from the visual of Univariate Analysis are as follows:

- Easy to identify the highest car sales city.
- Easy to compare the sales among various cities.
- Easy to understand the car sales trend across different cities.

These findings are important in terms of understanding the sales distribution across different cities and can be used to inform marketing and sales strategies.

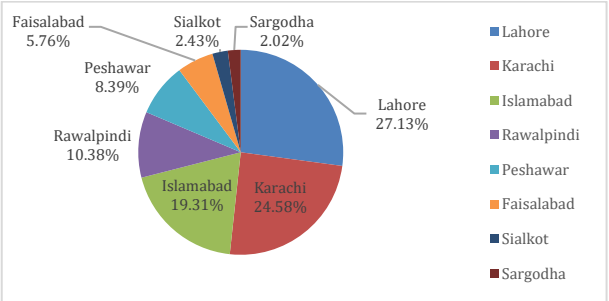

Fig. 4: Top cities with the most sales

The information about distribution and modeling type is necessary for deciding how to manage the variable in further analysis, such as which statistical methods to use. It is also useful in identifying any issues with the data, such as missing or extreme values that may need to be addressed before proceeding to other types of analysis.

One of the objectives of conducting univariate analysis is to gain an understanding of the dataset. For example, the pie chart was used to visualize the distribution of car models. We can collect information about the proportion of the top ten models of cars sold. The bar chart is used to visualize the distribution of car manufacturers. It shows the number of cars sold by the top ten car manufacturers. The map chart was used to visualize the distribution of sales locations. It shows the number of cars sold in various cities of Pakistan. This visual helps us to understand the data distribution and allows us to understand the market trends and consumer behavior in the automotive industry in Pakistan.

3.5. Bivariate analysis

It is used to analyze the relationship between two variables. It is conducted by data analysts to determine whether they are related and, if so, how they are related. The objective of bivariate analysis is to uncover patterns, trends, and relationships between two variables to gain a deeper understanding of the data. This technique can help to identify variables that are strong predictors of an outcome and can also provide insight into potential causal relationships between variables. Bivariate analysis is an important step in the data exploration process, as it allows researchers to explore the relationships between variables in a data set systematically.

The scatter matrix visualization in Fig. 5 highlights the relationship between car models and their respective prices using multiple perspectives. The top-left plot in Fig. 5 shows a diagonal line of points along the 45-degree axis, which serves as a reference indicating consistent values of price. The bottom-right scatter plot in Fig. 5 demonstrates how

prices vary across different car models. The upward trend of points from left to right suggests a positive relationship, meaning that as we progress through the sequence of models, their prices tend to rise. For instance, 'Alto' appears at the lower end of the price scale, 'Prado' occupies the highest price point around the middle, and models such as 'Mehran' return to the lower range. This indicates that higher-priced models are concentrated near the middle of the sequence, while lower-priced models appear at both extremes.

The top-right vertical bar plot and the bottom-left horizontal bar plot in Fig. 5 provide complementary views of the same information. They clearly illustrate the distribution of prices across models, with taller bars representing more expensive cars such as 'Prado' and shorter bars highlighting more affordable options such as 'Alto' and 'Mehran'. Together, these views reinforce the observation that the price distribution of car models is uneven, with a few high-priced models standing out against a majority of lower-priced ones.

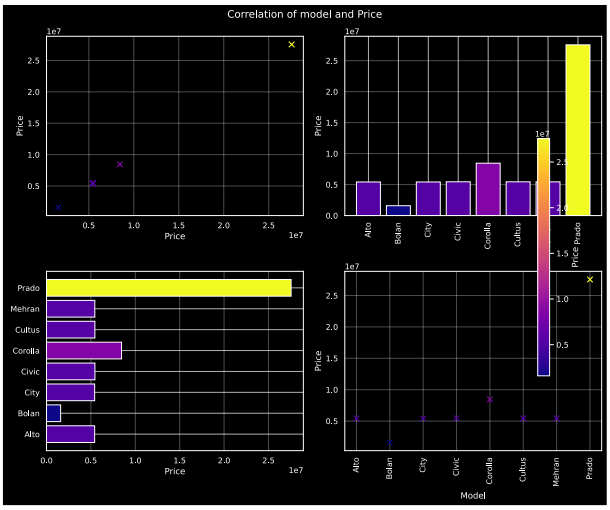

Fig. 5: Bivariate analysis, the relationship between price and model

3.6. Data preprocessing for predictive modelling

Data preprocessing is a crucial step in the data analysis and modeling process. It involves transforming raw data into a clean and structured format suitable for further analysis and modeling. The objective of preprocessing is to make the data usable and meaningful by removing or replacing missing or irrelevant values, handling outliers, transforming variables, and scaling data to meet the required format. Preprocessing makes the data suitable for various statistical analyses and ML algorithms by improving the accuracy, performance, and reliability of the models.

It is a critical step as it can greatly affect the results of the analysis and modeling. For example, if the data has outliers, these outliers may affect the mean and standard deviation, which will then have an impact on the results of the analysis. Similarly, if the data is not in the proper format, the algorithms may not work effectively. Therefore, data preprocessing is essential for ensuring that the data is ready for analysis and modeling.

The preprocessing process normally includes the following steps: data cleaning, data transformation, data scaling, and data normalization. The specific steps involved in the preprocessing process depend on the data, the objectives of the analysis, and the data preparation required for specific algorithms. In some cases, additional steps may be required, such as feature engineering, which involves creating new variables or features based on existing variables in the data set.

In this process, the first step was to drop the columns postedFrom, adLastUpdated, and features as they were deemed irrelevant to the analysis.

Then, the categorical variables in the data were encoded using LabelEncoder. This is because many statistical models require the input data to be numerical. Label Encoder is a technique used to convert categorical data into numerical data. It maps each unique categorical value to a unique integer value. The label encoder works by first finding all the unique categories in the categorical column and then encoding each category to an integer value. The resulting encoded values will be in the range of 0 to (number of unique categories - 1). The label encoder is useful in converting categorical data into numerical form so that it can be used as input in machine learning algorithms that require numerical data.

The categorical variables in the data included model, manufacturer, fuelType, vehicleTransmission, color, bodyType, RegisteredIn, Assembly, and location.

Finally, the numerical variables modelDate, mileageFromOdometer, and EngineCapacity were standardized by subtracting the mean and dividing by the standard deviation. This helps to ensure that all variables are on the same scale, making it easier to compare their effects on the outcome.

Let X be the numerical column with n data points. The meaning of X can be calculated as follows:

$$mean(x) = \frac{1}{n} \times \sum_{i=0}^{n} x_i \quad (1)$$

The standard deviation of X can be calculated as follows:

$$std(x) = \sqrt{(\frac{1}{n} \times \sum_{i=0}^{n} (x_i - mean(x))^2)} \quad (2)$$

The numerical columns in the data frame are transformed by subtracting the mean of the column and dividing by the standard deviation of the column. This normalization ensures that data is kept on a common scale, which is essential for the proper functioning of certain machine learning algorithms. The mathematical form of this transformation is given by:

$$x_{norm} = x - \frac{mean(x)}{std(x)} \quad (3)$$

3.7. Training the model

The next step after preprocessing the data is to train a predictive model. In this case, the XGBoost Regressor Model was selected for the task. XGBoost is a powerful and widely used machine learning algorithm that is specifically designed for gradient boosting decision trees. It is known for its fast performance, parallel processing capabilities, and high accuracy.

The model was trained on the processed data by providing it with the input features and the target variable. The training process involves iteratively updating the model parameters so that the predicted values come closer to the actual values. This process is repeated until a satisfactory level of accuracy is achieved. The XGBoost algorithm uses an ensemble of decision trees, each of which makes a prediction, and the final prediction is obtained by combining these predictions.

XGBoost Regressor is an implementation of gradient boosting decision trees. The algorithm works by building a series of decision trees and combining them to form an ensemble that predicts the target variable. The following are high-level mathematical equations that describe the training process in XGBoost Regressor:

Initial prediction: XGBoost Regressor starts with an initial prediction, often the mean of the target variable.

$$\hat{y}^{(0)} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad (4)$$

where, $\hat{y}^{(0)}$ is the initial prediction. y_i is the actual target value for instance i. n is the number of observations.

Loss function: The algorithm then uses a loss function to measure the difference between the actual and predictive variable. Common loss functions used in XGBoost Regressor include mean squared error (MSE) and mean absolute error (MAE).

MSE:

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

MAE:

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

Gradient calculation: The gradient of the loss function concerning the prediction is calculated to determine the direction in which the prediction should be adjusted to minimize the loss. If the mean squared error (MSE) loss function is used, it can be represented as:

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

The gradient of this loss function with respect to the prediction is:

$$\frac{\partial L}{\partial \hat{y}} = -\frac{2}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \quad (8)$$

Tree construction: The gradient information is then used to construct a decision tree that makes predictions based on the input variables. The tree is grown by repeatedly splitting the data into smaller subsets based on the values of the input variables. The tree construction process in XGBoost Regressor is designed to find the splits that best reduce the loss function and improve the prediction accuracy.

Ensemble formation: The process of tree construction is repeated multiple times to form an ensemble of decision trees. The final prediction is made by combining the predictions of all the trees in the ensemble.

In summary, XGBoost Regressor trains an ensemble of decision trees to make predictions by minimizing a loss function through gradient descent.

The training process also involves hyperparameter tuning, which is the process of selecting the best values for the parameters that control the model's behavior. This is being done to optimize the model's performance on the training data and ensure that it generalizes well to new data. The XGBoost algorithm has several hyperparameters that can be adjusted, including the learning rate, the number of trees, the depth of the trees, and the type of loss function used.

In conclusion, the training process for the XGBoost Regressor Model involves iteratively updating the model parameters to minimize the prediction error, as well as hyperparameter tuning to optimize the model's performance. The end goal is to build a model that can accurately predict the target variable based on the input features.

We used the XGBoost regressor, a powerful machine learning algorithm, to build predictive models. XGBoost is a gradient boosting framework that has proven to be highly effective for solving regression problems. In this study, we used the XGBoost regressor to build models that could predict the prices of automobiles based on their features.

3.8. Training results

Fig. 6 shows the graph of training results. Two curves can be observed: one for the training score and the other for the display validation score. The learning curve is a common visual sign used in ML to evaluate performance model training. The x-axis shows the number of training instances, and the y-axis shows the accuracy.

The learning curve depicts how the model's performance changes as the number of training examples increases and helps to identify problems of overfitting or underfitting. An efficient model will usually have the curve start low and gradually increase as the number of training examples increases. If the curve looks flat, it may indicate a problem with the model's capability to learn. Whereas a rapidly decreasing learning curve may indicate overfitting.

Another way to visualize the performance of the XGBoost Regressor is to use feature importance

plots. These plots help to identify the most important features used by the model to make predictions. By analyzing these plots, it is possible to gain sufficient information about the underlying patterns and relationships within the data. In addition, feature importance plots can be used to identify potential areas for further investigation and analysis. Another useful visualization is the confusion matrix, which is a table that is used to evaluate the performance of a classification model. It provides a summary of the number of correct and incorrect predictions made by the model for each class. By analyzing the confusion matrix, it is possible to gain an adequate understanding of the types of errors made by the model, such as false positives and false negatives. These visualizations can help to identify areas where the model may be improved and provide information for further research.

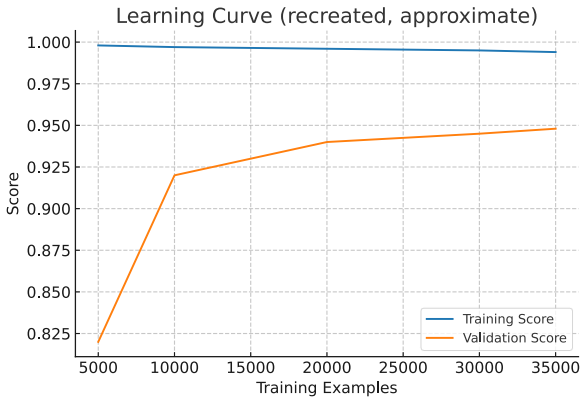

Fig. 6: Training and validation

In XGBoost Regressor, feature importance is a measure of how much each feature contributes to the final prediction. The feature importance can be visualized in a bar chart Fig. 7 where each horizontal bar represents the relative importance of each feature. In the visualization, it is evident that the most important feature is the model of the vehicle. This means that the model of the vehicle has the highest impact on the final prediction.

This information can be used to determine which features to include in the model and which features to eliminate. Additionally, this information can be used to gain a deeper understanding of the underlying relationships between the features and the target variable.

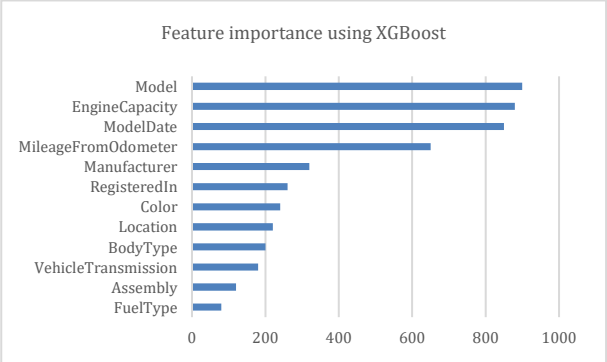

Fig. 7: Feature importance of the XGBoost regressor model

It is important to note that feature importance should not be used as the sole criterion for feature selection. It should be considered in conjunction with other metrics such as feature correlation, model performance, and domain knowledge.

4. Results

The outcomes of this research study reveal that AI tools are very effective for accurate anticipation of business sales volume. Business organizations can also approach market intelligence by using advanced analytics techniques. Business managers can conduct customer behavior studies with more detail. They may improve their sales forecasting capabilities. Which in turn enables them to compete in the market. They can plan business activities in a more effective way.

With the emergence of big data analytics, a huge amount of information is ready and available to business organizations. Big data is now proven to be an asset rather than useless garbage to dump. However, there is an immense need to utilize big data analytics with appropriate ML techniques.

With the rise of big data, businesses now have access to vast amounts of information that can be leveraged to gain awareness of their customers, the market, and their operations. However, traditional forecasting methods may not provide sufficient accuracy in the rapidly changing business environment.

We explore the potential of predictive AI in the field of sales forecasting. The study uncovers hidden patterns and makes informed predictions about future sales. The results of the study show a significant improvement in forecasting accuracy.

In this study, we developed an AI model based on XGBoost. The model was trained and tested to forecast future sales. The results proved that the model achieved an R2 score of 0.94. It indicates a higher predictive capability. The cross-validation results are also very encouraging, with a mean score of 0.94 and a standard deviation of 0.02.

These results show the potential for businesses to utilize big data analytics and predictive AI techniques. Business managers could gain a valuable understanding and make informed decisions. By incorporating big data analytics and predictive AI, businesses can increase their competitiveness in the market.

5. Discussion

We thoroughly review the prior research relevant to the objectives of our study. We aim to discover new avenues of predictive Artificial Intelligence in sales projection. It shows effective utilization of an AI tool supportive of business managers in smart decision-making. The results reveal that the use of a predictive AI model based on XGBoost improved the accuracy of sales forecasting. Achieved R2 score of 0.94. The cross-validation results also showed higher

stability with a mean score of 0.94 and a standard deviation of 0.02.

The results of this study further substantiate the importance of data analytics and its support to businesses. These are also in line with other studies that have shown the advantages of using big data and advanced analytics. There are many business aspects, such as sales forecasting, customer behavior analysis, and marketing optimization, that can be improved with the help of AI technologies. These findings also indicate the potential for businesses to obtain a competitive edge and make better decisions.

The findings of this research show the importance of the appropriate selection of ML technologies. We use the most suitable model, based on XGBoost shows high accuracy and stability. The results determine the future direction and explore new avenues for further studies.

List of abbreviations

AI	Artificial intelligence
ANN	Artificial neural network
CNN	Convolutional neural network
DEA	Data envelopment analysis
EDA	Explanatory data analysis
ERP	Enterprise resource planning
GRU	Gated recurrent unit
MAE	Mean absolute error
ML	Machine learning
MSE	Mean squared error
ORCID	Open Researcher and Contributor ID
RF	Random forest
R^2	R-squared (coefficient of determination)
SD	Standard deviation
SVM	Support vector machine
XGBoost	Extreme gradient boosting

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Ahaggach H, Abrouk L, and Lebon E (2024). Systematic mapping study of sales forecasting: Methods, trends, and future directions. Forecasting, 6(3): 502-532. https://doi.org/10.3390/forecast6030028

Ayub N, Irfan M, Awais M, Ali U, Ali T, Hamdi M, Alghamdi A, and Muhammad F (2020). Big data analytics for short and medium-term electricity load forecasting using an AI techniques ensembler. Energies, 13(19): 5193. https://doi.org/10.3390/en13195193

Cadavid JPU, Lamouri S, and Grabot B (2018). Trends in machine learning applied to demand and sales forecasting: A review. In the 7th International Conference on Information Systems, Logistics and Supply Chain, Lyon, France.

Caglayan N, Satoglu SI, and Kapukaya EN (2020). Sales forecasting by artificial neural networks for the apparel retail chain stores. In the Intelligent and Fuzzy Techniques in Big Data Analytics and Decision Making: Proceedings of the INFUS 2019 Conference, Springer International Publishing, Istanbul, Turkey: 451-456. https://doi.org/10.1007/978-3-030-23756-1_56

Faheem M, Aslam M, and Kakolu S (2024). Enhancing financial forecasting accuracy through AI-driven predictive analytics models. IRE Journals, 4(12): 322-328.

Fathima F, Inparaj R, Thuvarakan D, Wickramarachchi R, and Fernando I (2024). Impact of AI-based predictive analytics on demand forecasting in ERP systems: A systematic literature review. In the International Research Conference on Smart Computing and Systems Engineering, IEEE, Colombo, Sri Lanka, 7: 1-6. https://doi.org/10.1109/SCSE61872.2024.10550480

Gupta S, Chen H, Hazen BT, Kaur S, and Gonzalez ED (2019). Circular economy and big data analytics: A stakeholder perspective. Technological Forecasting and Social Change, 144: 466-474. https://doi.org/10.1016/j.techfore.2018.06.030

Iyengar NCS, Vivekanandan T, and Ahmed ST (2023). Big data analytic approach to predict risk assessment for cardiovascular diseases using Framingham risk score. International Journal of Computational Learning and Intelligence, 2(1): 32-38.

Kaggle (2024). Pakistan's largest Pakwheels automobiles listings. Available online at: https://www.kaggle.com/datasets/asimzahid/pakistans-largest-pakwheels-automobiles-listings

Okeleke PA, Ajiga D, Folorunsho SO, and Ezeigweneme C (2024). Predictive analytics for market trends using AI: A study in consumer behavior. International Journal of Engineering Research Updates, 7(1): 36-49. https://doi.org/10.53430/ijeru.2024.7.1.0032

Pavlyshenko BM (2019). Machine-learning models for sales time series forecasting. Data, 4(1): 15. https://doi.org/10.3390/data4010015

Punia S and Shankar S (2022). Predictive analytics for demand forecasting: A deep learning-based decision support system. Knowledge-Based Systems, 258: 109956. https://doi.org/10.1016/j.knosys.2022.109956

Rashidi SF, Olfati M, Mirjalili S, Platoš J, and Snášel V (2025). A comprehensive DEA-based framework for evaluating sustainability and efficiency of vehicle types: Integrating undesirable inputs and social-environmental indicators. Cleaner Engineering and Technology, 27: 100989. https://doi.org/10.1016/j.clet.2025.100989

Seyedan M and Mafakheri F (2020). Predictive big data analytics for supply chain demand forecasting: Methods, applications, and research opportunities. Journal of Big Data, 7(1): 53. https://doi.org/10.1186/s40537-020-00329-2

Smith MA and Côté MJ (2022). Predictive analytics improves sales forecasts for a pop-up retailer. INFORMS Journal on Applied Analytics, 52(4): 379-389. https://doi.org/10.1287/inte.2022.1119

Yun C, Shun M, Junta U, and Browndi I (2022). Predictive analytics: A survey, trends, applications, opportunities' and challenges for smart city planning. International Journal of Computer Science and Information Technology, 23(56): 226-231.