Journal of Data Analysis and Information Processing, 2024, 12, 510-522
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

AutoGluon-Based Sales Forecasting a Real-Time Predictive Analytics Solution for Business Intelligence

Fahamgeer Mahesar[1*], Anam Ishaq[1], Malaika Riaz[1], Muhammad Bilal[2], Tanzeela Kousar[3], Aqsa Jameel[3]

[1]Faculty of Information Technology, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, Pakistan
[2]Torch Solutions, Lahore, Pakistan
[3]Institute of Computer Science & IT, The Women University, Multan, Pakistan
Email: *fahamgirmahesar@gmail.com

How to cite this paper: Mahesar, F., Ishaq, A., Riaz, M., Bilal, M., Kousar, T. and Jameel, A. (2024) AutoGluon-Based Sales Forecasting a Real-Time Predictive Analytics Solution for Business Intelligence. *Journal of Data Analysis and Information Processing*, **12**, 510-522.
https://doi.org/10.4236/jdaip.2024.124027

Received: August 2, 2024
Accepted: September 23, 2024
Published: September 26, 2024

Abstract

Accurate sales forecasting is essential in the fast-paced world of business for effective strategic planning and resource allocation. However, traditional forecasting methods often lack precision and flexibility. This study aims to address this issue by incorporating machine learning (ML) techniques to improve forecasting accuracy and responsiveness to market changes. The methodology involves gathering extensive sales data and carefully preprocessing it to ensure quality. Various ML algorithms, such as time series analysis, regression models, and neural networks, are utilized to account for the complex and non-linear nature of sales patterns. These models are trained and validated using historical sales data, taking into consideration external factors like economic indicators and consumer trends. The results show a significant enhancement in forecast accuracy compared to traditional methods. The ML models effectively capture underlying trends and seasonal variations, providing reliable predictions that closely match actual sales results. Additionally, the models demonstrate strong adaptability, quickly adjusting to unexpected market shifts.

Keywords

Machine Learning, AutoGluon, Sales

1. Introduction

Sales forecasting is super important but can be pretty tough for businesses in the

tech industry. It's crucial for planning everything: marketing, client service, and finances. Getting a good sales forecast model isn't easy! There are some real challenges For instance, over forecasting can bump up costs and create excess supply, while under-forecasting might mean missing out on sales opportunities and disappointing customers [1]. Accurate forecasts can actually improve customer satisfaction, strengthen channel relationships, save a lot of money.

Many methods have been used for forecasting, like Back Propagation Neural Networks (BPNs). These are popular because they can identify patterns in data. However, BPNs often have issues with controlling many parameters and can struggle with overfitting their models. That's why Support Vector Regression (SVR) comes into play! It nicely handles nonlinear regression tasks and has shown to give better predictions than BPNs because it effectively captures trends from the data. SVR is frequently employed for forecasting things like sales flow, financial trends, even wind speeds. But watch out—when there are a lot of possible independent variables, SVR can become less reliable.

To tackle this challenge, Multivariate Adaptive Regression Splines (MARS) offer a great alternative for dealing with complex nonlinear issues. MARS shines when working with large datasets, especially in areas like electricity price forecasting or credit scoring [2]. Sales forecasting plays a key role in helping businesses plan ahead and gain an edge over competitors. Traditional time series models mostly focus on linear data and often miss the more complex nonlinear stuff [3].

Luckily, researchers are turning to soft computing methods to solve these nonlinear challenges. They're applying techniques such as fuzzy neural networks and evolutionary algorithms for better sales predictions. Several algorithms aim to address issues similar to ARIMA models that can interpret historical data points quite quickly [4]. However, these models can sometimes struggle with complex data patterns found in sales.

Even though ANN-based algorithms could help here, they require quite a bit of time to improve their prediction accuracy. This is where Extreme Learning Machine (ELM) algorithms come in handy! They offer quicker learning times while per-forming better than traditional algorithms. ELM reduces the hassle that comes with traditional learning methods like setting learning rates or stopping criteria and helps simplify everything. This also means it's great for real-time operations like control systems [5].

Now, back to sales forecasting! It's really essential for supply chain management and interactions between retailers and manufacturers. Manufacturers need solid forecasts for product planning while retailers want accurate sales predictions to guide purchasing decisions and minimize costs. Depending on what kind of business you're running, you might use human input, basic statistical models, or a mix of both for your sales forecasts.

This research is all about exploring the AutoGluon model for forecasting. It looks at past sales to help make future predictions smoother. The goal here is to check out different forecasting methods used in finance and see how well

AutoGluon performs with its chosen dataset. AutoGluon provides an automated machine learning approach that delivers regression models—making the forecasting process easier with little human effort needed!

Challenges with Conventional Approaches

Data Accuracy: Conventional methods may use outdated or unreliable sales data. Machine learning techniques overcome this issue by analyzing data more thoroughly. **Assumptions of Static Variables:** Conventional models assume that sales variables stay the same over time. Machine learning models adjust dynamically to changing circumstances.

The Role of Machine Learning

AutoGluon utilizes historical sales data to analyze patterns, seasonality, and trends. It offers model flexibility by automatically selecting and tuning machine learning algorithms like Linear Regression, Decision Trees, and Neural Networks to adapt to different sales conditions, as shown in Figure 1. The strategic advantages of AutoGluon-driven forecasts include improved inventory management, resource allocation, and revenue planning.

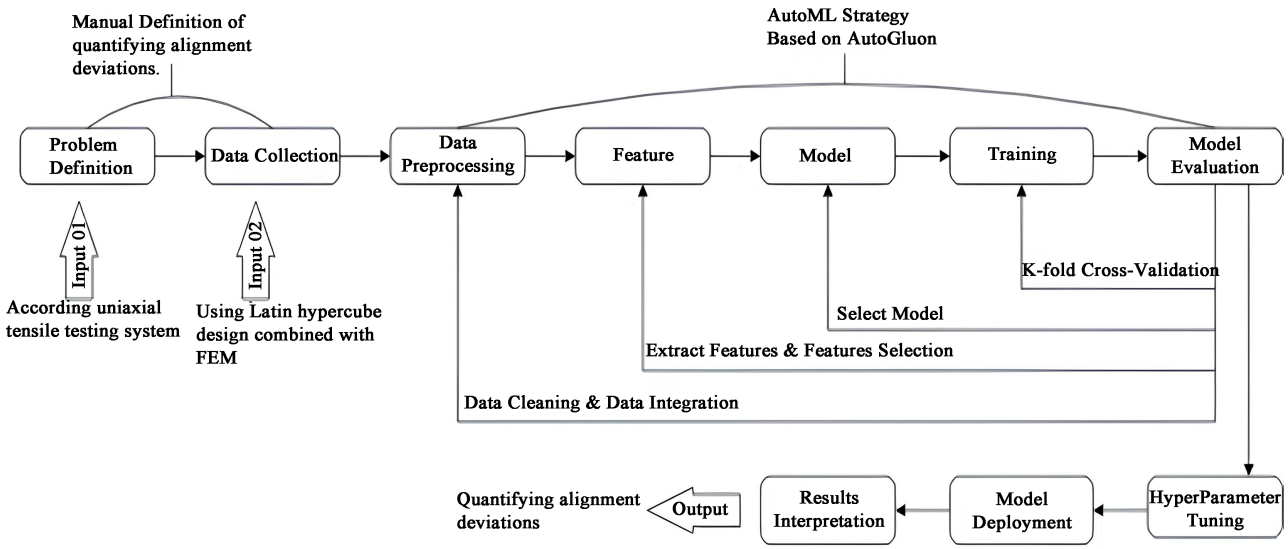

Figure 1. Research framework

2. Literature Review

Background: Supply Chain Management: Planning and Execution Supply chain management encompasses multiple business entities that collaborate on the movement of physical goods, customer service, and financial transactions. This field can be divided into two main areas: supply chain planning and supply chain execution.

Forecasting Concept: The concept of forecasting involves predicting future events, with sales forecasting being a projection of expected demand based on environmental conditions. Unlike predictable natural events, business forecasts can be prone to errors due to changing business equations over time. It's important to

note that forecasting is not the same as planning, as planning involves managerial actions to meet or exceed sales forecasts. The goal of accurate forecasting is to predict demand perfectly, and it is used across companies, service sectors, and government organizations as an input to planning activities. [6]. summarizes the characteristics of sales forecasts as follows: Predictions are never completely accurate and mistakes are bound to happen. Short-term forecasts tend to be more precise than long-term forecasts because the margin of error is smaller in relation to the average. Overall forecasts are more reliable than individual forecasts, as the margin of error is reduced. When there are significant discrepancies in the supply chain, sales forecasts are more likely to be inaccurate.

Sales Forecasting in Planning

Manufacturing industries strive to meet customer demand by providing the right supply [7]. emphasize the importance of sales forecasting in this process. End customers drive demand, which can be influenced by promotions. Marketing targets end customers to create demand, while the sales department works with wholesalers and retailers. Effective supply management requires collaboration between manufacturing, purchasing, and logistics to meet demand.

Forecasting Methods and Techniques

There are various standardized forecasting methods available, each with different levels of quantitative sophistication and logic base (such as historical data, expert opinion, or surveys). These methods can be grouped into historical projection, qualitative, and causal categories [8].

According to [8], projecting historical data into the future can effectively forecast short-term trends when a sufficient amount of data is available. Mathematical and statistical models are useful for these forecasts, especially in stable environmental conditions where demand patterns remain consistent year-to-year [7]. point out that different products may require different time series techniques to analyze historical sales patterns and make forecasts. Common time series patterns include level, trend, seasonality, and noise.

Machine Learning Techniques

There are three primary machine learning algorithms: supervised, unsupervised, and reinforcement learning. Supervised learning involves using labeled data to train models with known input-output relationships. Common supervised learning techniques include random forest, linear regression, and long short-term memory (LSTM) [9]. Unsupervised learning involves training models using unlabeled data. Techniques include association rule mining and clustering [10]. Association rule mining is the process of identifying connections between various items, such as patterns in supermarket transactions [11]. Clustering, on the other hand, involves categorizing data points into groups with similar characteristics [12].

3. Related Work

The field of sales forecasting is broad, with numerous approaches and methodologies

tailored to different contexts and data characteristics. Ongoing research is likely to continue focusing on improving model accuracy, handling complex data structures, and developing scalable solutions that can be applied in various industries. In their study, [13] utilized machine learning algorithms such as linear regression, random forest regression, and various time series techniques to forecast sales using Walmart's publicly available online sales data. Their models were developed using Azure Machine Learning Studio and R programming, with results showing that regression techniques were more effective than time series analysis. Alternatively, AutoGluon can simplify and expedite the modeling process. AutoGluon is an open-source AutoML toolkit that enables the creation of highly precise machine-learning models with minimal input.

AutoGluon for Sales Forecasting

AutoGluon streamlines the process of selecting and training models, simplifying the task of achieving accurate forecasts. By leveraging AutoGluon, we can combine various models and automatically identify the most effective ones. This method is capable of handling intricate, non-linear data patterns and can adapt to diverse forecasting requirements. Steps to implement AutoGluon include Data Preparation: Loading historical sales data, incorporating features such as product type, sales figures, promotions, etc. Model Training: Utilize AutoGluon to train multiple models, encompassing regression and time series techniques. Model Evaluation: AutoGluon assesses the performance of different models and chooses the optimal one based on predefined metrics. Deployment: Implement the top-performing model for real-time sales forecasting and integrate it into the supply chain management system. AutoGluon simplifies the process, minimizes the need for extensive parameter adjustments, and delivers reliable sales forecasts to facilitate supply chain planning and execution.

4. Methodology

To begin, we reviewed existing literature to gather relevant research. This information was used as a basis for our analysis of retail sales through the application of machine learning methods. Our main goal in this study is to assess the effectiveness of machine learning models using AutoGluon with sales data from point-of-sale systems. AutoGluon streamlines the training and selection of the most effective models by incorporating various techniques, including linear regression, Random Forest regression, and Xtreme Boosting Regression. The complete methodology of our proposed solution utilizing AutoGluon is depicted in **Figure 2**.

Dataset description

This paper outlines the methodology used for a retail Point of Sale system implemented in a test set of —S— = 32 locations in early 2007. The Citadel Point of Sales system was utilized to record all sales transactions, and data was collected from various tables using SQL queries. The study included multiple stores with different items for sale, each with five stations. The transaction history of one

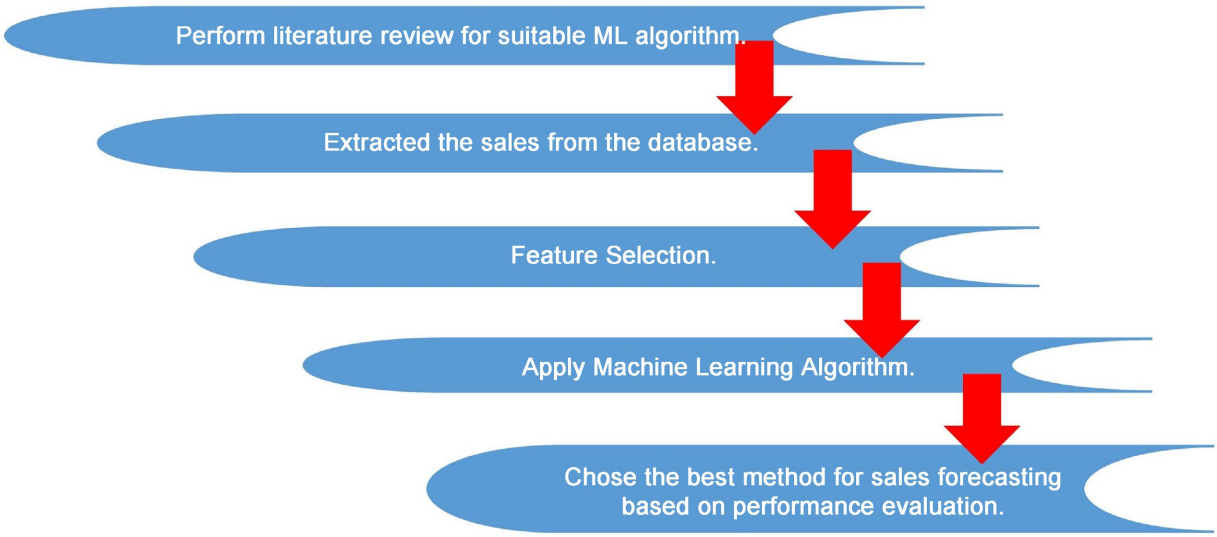

Figure 2. Sales forecasting flow diagram.

customer, who had 228 invoices, each containing an average of five items, was examined. Data from 2013 to 2018 was used for training, and testing was conducted on 2020 data. The training dataset consisted of item ID, store number, total sales items, and total sales of each item, totaling 87,847 rows.

Data Pre-Processing

We converted the data into days, weeks, and years, then identified and removed outliers, missing, or null values. The dataset was refined to perform testing. Several standardized methods for forecasting exist, differing in forecasting performance, the level of quantitative sophistication used, and the logic base from which the forecast is derived [8]. Augmented Dickey-Fuller Test This statistical test checks for the presence of a unit root in an autoregressive model. The null hypothesis (H0) is that a unit root is present, indicating the time series is non-stationary. The alternative hypothesis (H1) is that the time series is stationary [14]. P-value ¿ 0.05: Fail to reject H0, the data is non-stationary. P-value ¡ = 0.05: Reject H0, the data is stationary.

Specific Technique: We utilized the Augmented Dickey-Fuller (ADF) Test to determine the stationarity of our time series data due to its reliability and widespread use in detecting unit roots. Stationarity is crucial for accurate forecasting in models like ARIMA, and the ADF test is particularly effective as it accounts for higher-order autoregressive processes, offering more reliability than simpler tests. By providing a clear p-value-based decision criterion, the ADF test allows us to rigorously assess whether our data is stationary or requires further transformation. This choice ensures that our forecasting models are built on solid assumptions, thereby improving their accuracy and reliability.

Feature Selection

As there are many factors that play important role for machine learning success. Feature selection is a very important factor that has huge influence on machine

learning model performance. It helps from overfitting by removing data redundancy, reduces the training time and improves the accuracy of the model. There are different approaches which we used to overcome these problems, like correlation method. Feature sets having negative co-relations with target variables have been removed during the feature selection process.

Implementation

We utilized the AutoGluon model to analyze our results and compared its performance. Two metrics, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), were utilized to evaluate the effectiveness of the machine learning regression model. MAE is a common measure of forecast error in time series analysis and is used to assess the accuracy of predictions. It calculates the average of the absolute errors, with a lower error indicating higher accuracy of the model [15].

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{1}$$

where:

- y_i is the actual value;
- $\hat{y}_i$ is the predicted value;
- n is the number of samples.

Root Mean Square Error (RMSE) is the square root of the mean square error. It is the root of the average of squared differences between prediction and observation. Lower the error implies greater the accuracy of the model [15].

The Root Mean Square Error (RMSE) is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{2}$$

where:

- y_i is the actual value;
- $\hat{y}_i$ is the predicted value;
- n is the number of samples.

The regression equation takes the form,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon = \gamma \tag{3}$$

In the ever-changing world of business analytics, the use of machine learning algorithms for customer sales forecasting has become essential for organizations looking to improve their predictive capabilities and decision-making processes. With the abundance of data available, companies are turning to advanced analytical techniques to gain insights into customer behavior and market trends. This research paper explores various machine learning algorithms used in customer sales forecasting, comparing predictive models and techniques. By studying the impact of feature selection on the accuracy of sales forecasting models, the goal is to understand the factors that contribute to the effectiveness of predictive analytics in sales forecasting. Additionally, the paper examines key performance metrics used to evaluate the efficacy of predictive models, providing a comprehensive

overview of evaluation criteria necessary for assessing the predictive power and reliability of machine learning algorithms in customer sales forecasting. Through a systematic exploration of these topics, this research paper aims to provide valuable insights into the field of business analytics and guide decision-makers on best practices for using machine learning in sales forecasting strategies.

Different machine learning algorithms are utilized for customer sales forecasting to analyze historical sales data and identify patterns, trends, and relationships crucial for accurate predictions. Regression algorithms are popular for modeling relationships between variables and spotting sales patterns for forecasting purposes. Implementing machine learning models helps predict sales, identify benchmarks, analyze incremental impacts, and make informed decisions for the future. Various algorithms like time-series analysis, regression analysis, and decision trees are commonly used in customer sales forecasting to create robust predictive models that combine customer data with historical sales performance and external factors, providing more accurate predictions compared to traditional methods. By leveraging machine learning, businesses can better prepare for the future, navigate the dynamic market landscape, improve sales forecasting accuracy, and enhance overall business strategies.

Feature selection plays a crucial role in constructing accurate sales forecasting models by influencing the precision of multivariate regression models. Careful consideration of features during model training is essential to unveil patterns and trends that impact sales predictions. Selecting relevant features, such as sales representatives' performance and experience, can significantly enhance the accuracy of sales forecasting models. Feature selection acts as a linchpin in developing precise sales forecasting models, enabling businesses to make informed decisions based on reliable predictions.

To evaluate the effectiveness of predictive models in sales forecasting, key performance metrics must be considered. The reliability and relevance of data used in machine learning for sales forecasting are fundamental for assessing the effectiveness of predictive models. Monitoring benchmark performance is essential to ensure the efficacy of predictive models in sales forecasting. Establishing success metrics for predictive models in collaboration with the software development team and business owner before constructing a demand forecasting solution aligns expectations and objectives. Historical data is critical for analyzing past sales performance, generating accurate sales forecasts, and making informed decisions. Adjusting the sales strategy based on customer sentiment data can provide valuable insights into customer preferences and behaviors, serving as a key performance metric for evaluating predictive models in sales forecasting. Continuously adapting the sales strategy to meet customer needs is vital in assessing the effectiveness of predictive models in sales forecasting, reflecting the model's ability to respond to dynamic market conditions and customer demands.

Using machine learning techniques for customer sales forecasting is a promising approach for businesses looking to improve their predictive abilities and

decision-making processes. This research paper compares predictive models and techniques, emphasizing the importance of advanced algorithms like regression analysis, time-series analysis, and decision trees for extracting valuable insights from historical sales data. These machine learning algorithms can identify patterns, trends, and relationships within datasets to develop more accurate forecasting models compared to traditional methods. Feature selection is also crucial for improving the precision of sales forecasting models by addressing informational gaps and data disorder. The dynamic nature of market conditions and customer preferences requires continual adaptation of sales strategies, which is essential for evaluating the effectiveness of predictive models. Incorporating customer sentiment data into the analysis further enhances the predictive capabilities of these models, providing valuable insights into customer behaviors and preferences. Establishing success metrics in collaboration with stakeholders ensures that demand forecasting solutions meet the specific needs of the business. Overall, this research demonstrates the potential of machine learning in driving informed decision-making and helping businesses thrive in competitive markets.

5. Results and Discussion

Model Performance Metrics

In this section, we present the performance metrics of the AutoGluon model used for sales forecasting. The performance of the model is evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

Following Table 1 below summarizes the performance metrics of the AutoGluon model.

Table 1. Model performance using AutoGluon.

Metric	Value
RMSE	0.0470455
MAE	0.0216292
MSE	0.0022132

Explanation of Metrics

The Mean Absolute Error (MAE) measures the average magnitude of errors in a set of predictions, without considering their direction. It is calculated as the mean of the absolute differences between predicted and actual values, with each difference given equal importance. On the other hand, the Mean Squared Error (MSE) calculates the average of the squared errors, making it more sensitive to outliers compared to MAE. This is because MSE squares the errors before averaging, giving more weight to larger errors.

The Root Mean Squared Error (RMSE) is the square root of the average of squared differences between predicted and actual values. It offers a measure of the average error size, combining the units of the data being observed, which makes

it easier to interpret.

Model Results

The results indicate that the AutoGluon model has demonstrated (provide a brief summary of the model's performance, e.g., high accuracy, good generalization, etc.). The low values of MAE, MSE, and RMSE suggest that the model predictions are close to the actual values, indicating effective performance in forecasting sales.

Visual Representation of Model: Below is an image of the AutoGluon model, which helps to illustrate its structure and performance. This visual representation assists in understanding the different layers and components utilized in the forecasting process.

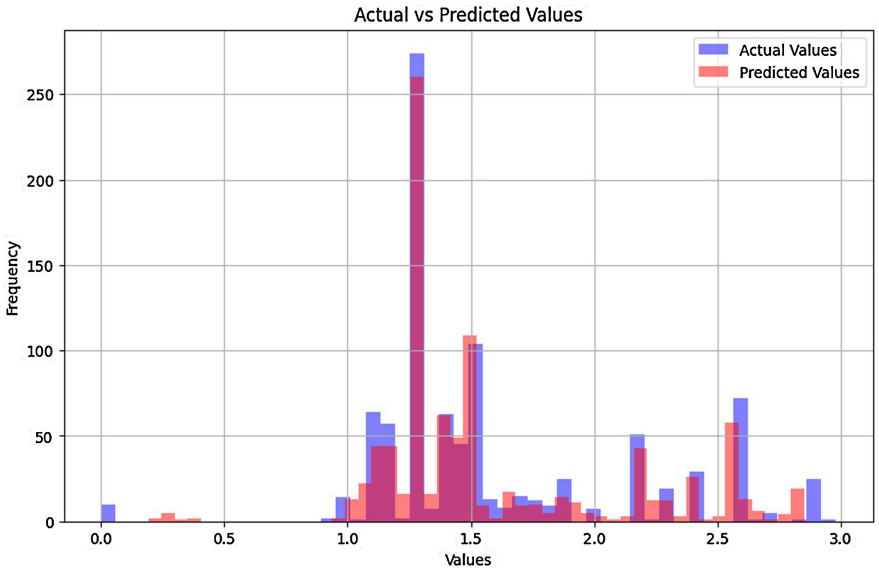

Figure 3. AutoGluon model's prediction.

In Figure 3, the histogram shows that while the predicted values closely follow the distribution of the actual values, there are areas where the model may be slightly off, either overestimating or underestimating the actual values. The overall alignment suggests a reasonable predictive performance of the model.

In Figure 4, the WeightedEnsemble-L2 model has established itself as the top-performing model for prediction within the AutoGluon models. It attained the highest validation score, demonstrating its superior accuracy and reliability.

Discussion

The results obtained from the AutoGluon model illustrate its capability in accurately predicting sales data. The low error metrics (MAE, MSE, and RMSE) reflect the model's precision in capturing the patterns in the historical sales data.

Several factors contribute to the model's effectiveness:

Automated Feature Engineering: AutoGluon automatically generates new features that enhance the model's predictive power, capturing intricate relationships within the data.

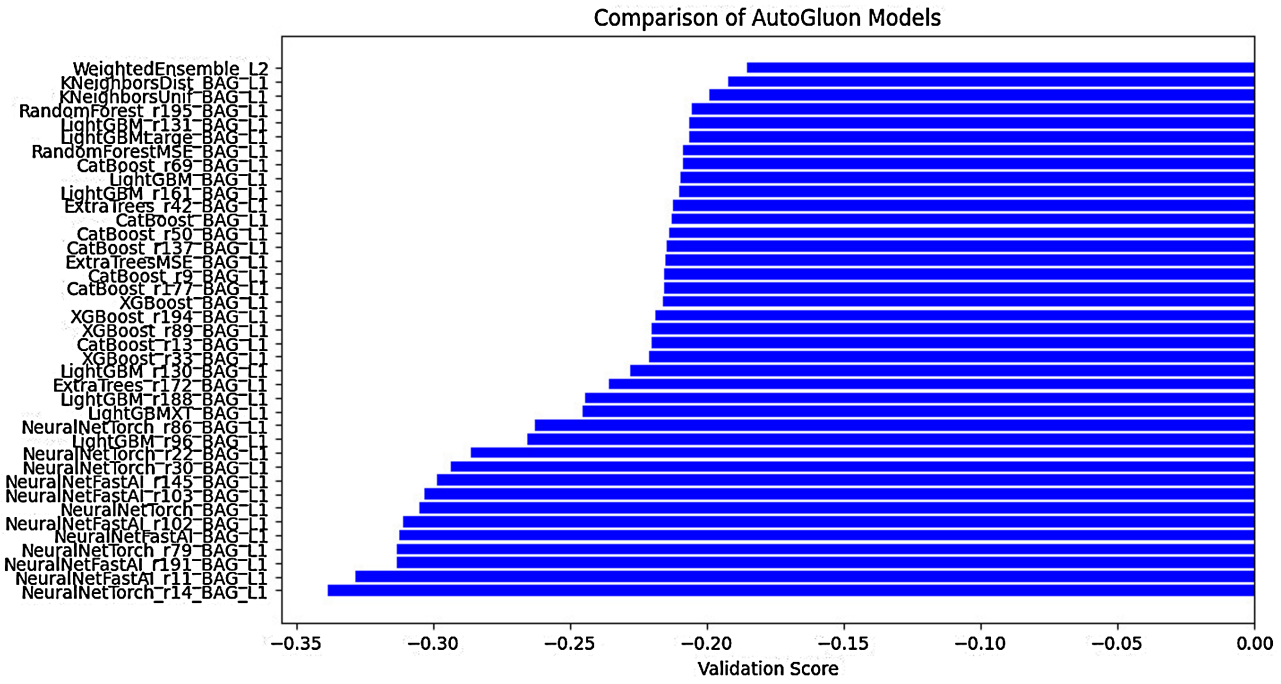

Figure 4. AutoGluon model's comparison.

Ensemble Learning: The use of ensemble methods in AutoGluon combines the strengths of multiple models, leading to improved overall performance.

Hyperparameter Optimization: AutoGluon's built-in hyperparameter optimization ensures that the best possible model configurations are selected, further enhancing accuracy.

6. Conclusion

In our research, we concluded that sales forecasting poses significant challenges for inventory management, marketing, customer service, and business financial planning within information technology chain stores. Accurate sales forecasting is crucial for supply chain management, bridging the gap between retailers and manufacturers. Manufacturers rely on predicting future demand to inform production planning, while retailers need accurate sales forecasts for purchasing decisions and cost optimization. Sales forecasting can be approached through human planning, statistical models, or a combination of both. Developing an accurate forecasting model is difficult due to factors like over-forecasting and under-forecasting. However, robust and precise sales forecasts yield benefits such as customer satisfaction, improved channel relationships, and substantial cost savings. In our study, we explored both time series models (such as LSTM) and machine learning regression algorithms using AutoGluon to predict sales. Notably, the AutoGluon model performed exceptionally well on the Citadel POS dataset. Looking ahead, deep learning approaches within AutoGluon can enhance sales forecasting by leveraging larger datasets. By applying deep learning models, we can further improve accuracy, especially when dealing with extensive retail sales data.

Acknowledgements

Authors would like to thank Dr. Maha Alabduljalil for her valuable feedback and ongoing encouragement.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Ofoegbu, K. (2021) A Comparison of Four Machine Learning Algorithms to Predict Product Sales in a Retail Store. Ph.D. Thesis, Dublin Business School.

[2] Deo, R.C., Kisi, O. and Singh, V.P. (2017) Drought Forecasting in Eastern Australia Using Multivariate Adaptive Regression Spline, Least Square Support Vector Machine and M5tree Model. *Atmospheric Research*, **184**, 149-175. https://doi.org/10.1016/j.atmosres.2016.10.004

[3] Álvarez-Díaz, M., González-Gómez, M. and Otero-Giráldez, M. (2018) Forecasting International Tourism Demand Using a Non-Linear Autoregressive Neural Network and Genetic Programming. *Forecasting*, **1**, 90-106. https://doi.org/10.3390/forecast1010007

[4] Shumway, R.H. and Stoffer, D.S. (2017) ARIMA Models. Time Series Analysis and Its Applications. Springer.

[5] Lu, C. and Kao, L. (2016) A Clustering-Based Sales Forecasting Scheme by Using Extreme Learning Machine and Ensembling Linkage Methods with Applications to Computer Server. *Engineering Applications of Artificial Intelligence*, **55**, 231-238. https://doi.org/10.1016/j.engappai.2016.06.015

[6] Hofmann, E. (2013) Supply Chain Management: Strategy, Planning and Operation. Elsevier Science.

[7] Mentzer, J. and Moon, M. (2005). Sales Forecasting Management: A Demand Management Approach. SAGE Publications. https://doi.org/10.4135/9781452204444

[8] Ballon, R. (2004) Business Logistics/Supply Chain Management: Planning, Organizing and Controlling the Supply Chain. 5th Edition, Pearson.

[9] Pavlyshenko, B. (2019) Machine-learning Models for Sales Time Series Forecasting. *Data*, **4**, Article 15. https://doi.org/10.3390/data4010015

[10] Hussain, S., Atallah, R., Kamsin, A. and Hazarika, J. (2018) Classification, Clustering and Association Rule Mining in Educational Datasets Using Data Mining Tools: A Case Study. In: Silhavy, R., Ed., *Cybernetics and Algorithms in Intelligent Systems*, Springer, 196-211. https://doi.org/10.1007/978-3-319-91192-2_21

[11] Kaur, M. and Kang, S. (2016) Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining. *Procedia Computer Science*, **85**, 78-85. https://doi.org/10.1016/j.procs.2016.05.180

[12] Sinaga, K.P. and Yang, M. (2020) Unsupervised K-Means Clustering Algorithm. *IEEE Access*, **8**, 80716-80727. https://doi.org/10.1109/access.2020.2988796

[13] Catal, C., Ece, K., Arslan, B. and Akbulut, A. (2019) Benchmarking of Regression Algorithms and Time Series Analysis Techniques for Sales Forecasting. *Balkan Journal of Electrical and Computer Engineering*, **7**, 20-26. https://doi.org/10.17694/bajece.494920

[14] Glynn, J., Perera, N. and Verma, R. (2007) Unit Root Tests and Structural Breaks: A

Survey with Applications.
https://www.researchgate.net/publication/30387415_Unit_Root_Tests_and_Structural_Breaks_A_Survey_with_Applications

[15] Chai, T. and Draxler, R.R. (2014) Root Mean Square Error (RMSE) or Mean Absolute Error (MAE). *Geoscientific Model Development Discussions*, **7**, 1525-1534.