WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A STUDY ON PREDICTING SALES USING ML ALGORITHMS AT SAI BABA AUTO COMPONENTS PRIVATE LIMITED

Santhosh R, Shireen Fathima S

PANIMALAR ENGINEERING COLLEGE

ABSTRACT :

This study explores the application of Machine Learning (ML) algorithms for sales forecasting, aiming to improve accuracy and efficiency compared to traditional methods. We investigate the use of Supervised Machine Learning algorithms to analyze historical sales data. The study emphasizes the potential of ML in sales forecasting and its contribution to data-driven decision making for businesses. It explores the impact of ML-powered sales forecasting on various aspects like inventory management, resource allocation, and targeted marketing campaigns. This research contributes to the understanding of how ML can enhance sales forecasting accuracy and its potential benefits for business performance.

Keywords : Order ID, Customer ID, SKU, Unit Price, MAE, MSE, RMSE, MAPE, R-squared, CPI, AI, SVR, IDE, GBM, Year, Month, Date, Quantity, HPLC

INTRODUCTION :

Accurate sales forecasting is paramount for businesses to achieve sustainable competitive advantage. Traditionally, this domain relied on a blend of subjective methods like sales experience and intuition, coupled with historical sales trend analysis. However, these approaches are inherently susceptible to human bias and struggle with the vast amount of data available in today's digital age. This limitation restricts their ability to identify complex relationships within datasets, potentially overlooking crucial factors influencing sales.The revolutionary paradigm shift is taking place, driven by the transformative power of Machine Learning (ML). This technology empowers us to harness the vast potential of historical sales data, uncovering hidden patterns, trends, and relationships that were previously invisible. By leveraging sophisticated algorithms and statistical models, ML facilitates the creation of highly accurate and dynamic sales forecasting models. This paper delves into the potential of ML to revolutionize sales forecasting, exploring its core principles, multifaceted approach, and transformative benefits for businesses. We meticulously construct robust ML models that analyze historical sales data encompassing a wide range of variables to gain a comprehensive understanding of the forces influencing sales performance. By implementing ML-powered sales forecasting, businesses can unlock a new era of data-driven decision making, optimized resource allocation, and unparalleled levels of sales success. This exploration contributes groundbreaking insights into the future of sales forecasting, fueled by the unparalleled power of Machine Learning.

NEED OF THE STUDY :

Traditional sales forecasting methods, reliant on experience and limited data analysis, struggle to keep pace with the dynamic market of today. Machine Learning (ML) offers a powerful alternative, leveraging sophisticated algorithms and vast historical sales data to uncover hidden patterns and trends crucial for accurate predictions. ML empowers businesses to move beyond basic forecasts, providing granular insights at the product, customer, and regional level. These insights, coupled with the ability to adapt to market shifts through real-time data integration, position businesses for unparalleled success. This project delves into the transformative potential of ML by developing robust models that analyze a wide range of sales-influencing factors. By implementing these ML-powered forecasts, businesses can unlock a new era of data-driven decision making, optimizing inventory management, resource allocation, marketing campaigns, pricing strategies, and proactive risk management. In essence, this project investigates the viability of ML for sales forecasting, aiming to contribute groundbreaking insights and empower businesses to achieve unparalleled levels of sales success.

OBJECTIVES OF THE STUDY :

1. To Train an SVR model using the company's historical sales data to assess its effectiveness in predicting future sales. This involves analyzing the model's accuracy and identifying potential areas for improvemenTo assess the risk management implications of integrating technical analysis into short-term trading strategies.
2. To Train a Random Forest model using the same sales data to compare its performance with the SVR model. Evaluating the accuracy of both algorithms will allow us to identify the best-suited model for the company's specific needs.

3. To Refine both SVR and Random Forest models to achieve the highest possible accuracy in sales prediction. This may involve techniques like hyperparameter tuning and feature engineering.
4. To Evaluate Based on the accuracy and interpretability of both SVR and Random Forest models, recommend the best performing model to the company for implementing in their sales forecasting process.

SCOPE OF THE STUDY :

This project, titled "Predicting Sales with Machine Learning Models," delves into the transformative potential of Machine Learning (ML) for generating accurate sales forecasts. Traditional methods, plagued by subjectivity and limited data analysis capabilities, are demonstrably inadequate in today's dynamic market.This study aims to leverage the power of ML to create a paradigm shift in sales forecasting. We will achieve this by constructing robust ML models that meticulously analyze vast historical sales data. These models will act as a powerful lens, uncovering hidden patterns and trends that were previously invisible and hold the key to accurate future sales predictions.The scope extends beyond basic forecasting. Our ML models will be designed to provide granular insights at a product, customer segment, and geographic region level. This comprehensive understanding of sales drivers empowers businesses to tailor strategies with laser focus, maximizing their effectiveness. Furthermore, the models will be equipped with the remarkable ability to continuously adapt to the ever-changing market landscape by integrating real-time data streams. This ensures that forecasts remain relevant and accurate, providing businesses with a crucial competitive edge.

LIMITATIONS OF THE STUDY :

- The accuracy of ML models hinges on the quality and quantity of training data. Limited company data or inconsistencies within sales records can restrict the model's ability to learn complex patterns and lead to unreliable forecasts.
- Biases present in the training data can be reflected in predictions, potentially skewing forecasts. Additionally, historical trends may not always predict future market shifts or disruptions not present in the data. Overly complex models can also become overly specialized, failing to adapt to unforeseen situations like new markets or products.
- Understanding the inner workings of complex ML models can be difficult. This limited interpretability might hinder trust and adoption, as stakeholders may struggle to understand or justify specific sales forecasts.
- The study acknowledges the limitations in predicting entirely unexpected events beyond the model's training data. These unforeseen external events, such as economic shifts, natural disasters, or major changes in consumer behavior, can significantly impact sales and require adjustments to the models.

REVIEW OF LITERATURE :

1. "Sales Forecasting with Machine Learning: A Review" by Nikhil et al. (2023**)** investigates the application of machine learning (ML) in sales forecasting. Their meta-analysis, a technique summarizing multiple research papers, reveals a significant trend: ML models outperform traditional forecasting methods in terms of accuracy. This translates to more reliable predictions of future sales for businesses leveraging ML's power. The study delves into specific ML techniques contributing to this improvement, providing valuable insights for businesses considering adopting these advanced forecasting methods.

2. *"Improving Sales Forecasting Accuracy with Machine Learning": A Comparative Analysis" by Wu et al. (2022)* focuses on a comparative analysis of various machine learning (ML) algorithms for sales forecasting. Their research emphasizes the potential of ML to surpass traditional methods, particularly when dealing with complex sales data. This suggests that ML can provide businesses with a significant advantage in forecasting future sales, especially when the data involves intricate patterns and relationships.

RESEARCH METHODOLOGY :

Research methodology refers to the overall approach, strategies, techniques, and processes used by researchers to systematically gather, analyze, and interpret information in their investigations. It encompasses the entire framework within which research is conducted, guiding how researchers plan, structure, and execute their studies. A well-defined methodology helps ensure that research is rigorous, valid, and reliable.

RESEARCH DESIGN :

Research design is the overall strategy or framework that guides a research project from its inception to its conclusion. It provides a systematic plan for collecting, analyzing, and interpreting data, ensuring that the study addresses its research questions effectively and reliably. A well-thought-out research design allows researchers to achieve their objectives while minimizing potential biases or errors.

DATA COLLECTION METHODS

Data was gathered from secondary sources. This process involves utilizing data previously collected by others for a different reason. Researchers examine and analyze this data to draw out useful information. Secondary data can come from multiple sources.

MACHINE LEARNING ALGORITHMS USED IN PYTHON AND ORANGE

RANDOM FOREST REGRESSION MODEL

Random Forest Regression is a supervised learning technique for continuous value prediction. It utilizes an ensemble approach, combining predictions from multiple decision trees to achieve a more robust and accurate final prediction.**:**

- **Decision Trees:** These flowchart-like structures split data based on features to make predictions. At each node, a decision rule based on a specific feature directs data down branches until reaching "leaf nodes" with the final prediction.
- **Ensemble Learning:** Random Forest builds diverse decision trees by introducing randomness during training. This can involve randomly selecting features or data points for each tree. By combining these diverse trees, the model reduces variance and avoids overfitting.
- **Training:** The process involves defining the number of trees to build and for each tree:
    - Randomly selecting a subset of features to consider at each split.
    - Randomly sampling a subset of data points (bootstrapping) to train the tree.

The final prediction for a new data point is the average of the predictions from all the trees in the forest.

GRADIENT BOOSTING REGRESSION MODEL

Gradient boosting regression utilizes a unique approach. It relies on a series of simple models (often decision trees) with limited individual predictive power, acting as building blocks for the final ensemble. The model works in stages, iteratively improving predictions. In each stage, the model calculates the difference between actual values and current predictions (residuals) and trains a new weak learner specifically to predict these errors. This allows the model to learn from past mistakes and focus on correcting them. Finally, the predictions from the new weak learner are added to the current prediction, accumulating knowledge and improving overall accuracy.

3.DECISION TREE REGRESSION MODEL

A decision tree regression model, a supervised learning technique for continuous value prediction, builds a tree-like structure to arrive at final predictions. It starts with the entire dataset at the root node and recursively splits the data based on features (attributes) using a splitting criterion (e.g., minimizing variance). At each split, a decision rule is created based on a specific feature to separate the data into groups with more similar target variable values. This process continues until a stopping criterion (e.g., maximum depth, minimum data points) is met. To make predictions, a new data point traverses the tree, following the decision rules at each node based on its feature values, until it reaches a final leaf node with the predicted value.

4.SUPPORT VECTOR REGRESSION MODEL

Support Vector Regression (SVR) is a supervised learning technique for predicting continuous numerical values. Unlike linear regression, which minimizes squared errors, SVR finds a hyperplane (a higher-dimensional line) that minimizes prediction errors while ensuring a margin of support around it. This margin is maximized to create a robust model.SVR separates data points based on their target variable values using a hyperplane. The data points closest to the hyperplane on either side are called support vectors, as they define the hyperplane's position and the margin. SVR uses an epsilon-insensitive loss function that penalizes only predictions deviating from actual values by more than a predefined threshold (epsilon). This allows for a certain tolerance for errors and focuses on fitting the data within the margin. SVR comes in two forms: linear SVR with a straight-line decision boundary, and non-linear SVR with kernel functions that project data into higher dimensions for effective separation using linear hyperplanes. This study utilized non-linear SVR for the regression model.

5.ARTIFICIAL NEURAL NETWORK REGRESSION MODEL

Neural networks, inspired by the human brain, are powerful tools for regression tasks like sales prediction. They consist of artificial neurons, the building blocks, that receive inputs, apply an activation function (introducing non-linearity), and generate outputs. These neurons are organized into layers: an input layer for raw features, hidden layers for processing information, and an output layer for the final prediction (sales amount in this case). A neural network learns by iteratively adjusting connections between neurons. Data flows through the network (forward pass), predictions are compared to actual values (loss calculation), and errors are propagated back (backpropagation) to adjust weights and minimize future loss. This process continues until the network can accurately predict sales based on features.This study utilizes a neural network with one hidden layer of 10 neurons. However, several hyperparameters can be tuned to improve performance:

- **Network Architecture:** Experimenting with different numbers of hidden layers and neurons can improve the model's ability to capture complex relationships without overfitting.

- **Activation Function:** Different activation functions (e.g., Leaky ReLU, Sigmoid, Tanh) can be explored to optimize the model's learning based on the data's characteristics.
- **Optimizer:** Tuning the learning rate or using alternative optimizers (SGD, RMSprop) can influence the efficiency of the learning process.
- **Regularization:** L1 or L2 regularization techniques can be applied to prevent overfitting if observed.
- **Training Iterations:** Adjusting the number of training iterations can be done based on the convergence rate and validation performance.

By carefully tuning these hyperparameters, the neural network's performance in predicting sales amounts can be further optimized.

SUMMARY OF FINDINGS :

EVALUATION METRICS

- **Mean Squared Error (MSE**): This metric measures the average squared difference between predicted and actual sales values. A lower MSE indicates a better fit, signifying smaller average discrepancies between predictions and actual sales. While it provides a quantitative measure, interpreting the raw MSE value can be challenging. High MSE values suggest larger overall discrepancies.
- **Root Mean Squared Error (RMSE):** As the square root of MSE, RMSE offers a more interpretable measure in the same units as the target variable (sales amount). A lower RMSE indicates a smaller average difference between predicted and actual sales values. In this case, an RMSE of 6703 suggests a significant average difference.
- **Mean Absolute Error (MAE):** Unlike MSE, MAE focuses on the average absolute difference between predicted and actual sales. This metric is less sensitive to outliers that can inflate MSE. A lower MAE indicates a smaller average prediction error in the same units as the target variable. Here, a value of 2997 suggests an average model deviation of around 2997 units from actual sales figures.
- **Mean Absolute Percentage Error (MAPE):** This metric calculates the average absolute percentage difference between predicted and actual sales amounts. It's valuable for comparing models on datasets with different scales because it expresses the error as a percentage. A MAPE value of 0.012 suggests an average error of only 1.2%. However, it's crucial to consider the range of sales figures, as a 1.2% error might translate to a larger absolute difference depending on the sales value.
- **R-squared (R^2):** This metric reflects the proportion of variance in the actual sales data that the model can explain. R^2 ranges from 0 to 1, with a higher value indicating a better fit. An R^2 value of 0.752 suggests that the model explains 75.2% of the variance observed in the actual sales data. It's important to remember that a high R^2 doesn't necessarily guarantee good prediction accuracy, especially for unseen data. Therefore, using R^2 in conjunction with other metrics like MSE, RMSE, and MAE provides a more comprehensive picture of model performance.

RANDOM FOREST REGRESSION MODEL EVALUATION

Model	MSE	RMSE	MAE	MAPE	R2
Random Forest	44936293	6703	2997	0.012	0.752

GRADIENT BOOSTING REGRESSION MODEL

Model	MSE	RMSE	MAE	MAPE	R2
Gradient Boosting	49291830	7020	5489	0.02	0.73

DECISION TREE REGRESSION MODEL

Model	MSE	RMSE	MAE	MAPE	R2
Decision Tree Regression	50027968	7073	3807	0.02	0.72

SUPPORT VECTOR REGRESSION MODEL

Model	MSE	RMSE	MAE	MAPE	R2
SVR Regression	68959533	8304	5648	0.03	0.71

ARTIFICIAL NEURAL NETWORK REGRESSION MODEL

Model	MSE	RMSE	MAE	MAPE	R2
ANN Regression	198440839	14086	11804	0.056	0.163

COMPARISON OF MACHINE LEARNING MODELS:

Model	MSE	RMSE	MAE	MAPE	R2
Random Forest	44936293	6703	2997	0.012	0.752
GradientBoosting	49291830	7020	5489	0.02	0.73
Decision Tree Regression	50027968	7073	3807	0.02	0.72
SVR	68959533	8304	5648	0.03	0.71
ANN Regression	198440839	14086	11804	0.056	0.163

FINDINGS:

- **Random Forest:** It has the lowest MSE, RMSE, and MAE, indicating the best performance in terms of minimizing prediction errors. It also has a relatively high R- squared value, suggesting it explains a good portion of the variance in the target variable.
- **Gradient Boosting:** It has a slightly higher MSE, RMSE, and MAE compared to Random Forest, but still lower than Decision Tree and SVR. Its R-squared is lower thanRandom Forest, indicating a potentially weaker fit.
- **Decision Tree Regression:** It has a higher MSE, RMSE, and MAE compared to Random Forest and Gradient Boosting. It also has a lower R-squared value, suggestingthe weakest fit among the four models.
- **SVR:** It has the higher MSE, RMSE, and MAE, along with the lowest R-squared value.It also has a lower R-squared value, suggesting the weakest fit among the four models.
- **Neural Network:** It has the highest MSE, RMSE, and MAE, along with the lowest R-squared value. This suggests the poorest performance in terms of minimizing errors andexplaining the data variance.
- In conclusion**,** based on this data, the Random Forest model appears to be the best performer for this task. It achieves the lowest prediction errors and explains a relatively high proportion of the variance in the target variable.

PREDICTING SALES FOR THE YEAR 2024 USING RANDOM FOREST:

The process of predicting sales of 2024 includes these steps:

1. **Data Augmentation**: We began by enriching the existing dataset with features that capture historical trends. This involved incorporating information from previous years'sales data.
2. **Growth Rate Calculation:** We calculated the average sales growth rate across the historical data. This provided a quantitative measure of how sales have typically changed year-over-year.
3. **Incorporating Variation**: To account for potential fluctuations in future sales, we introduced a controlled level of variation. In this case, we added a variation of 5.11% to the average growth rate. This variation could be derived from market analyses, seasonal trends, or other relevant factors.

4. **Feature Adjustment:** We then applied the calculated growth rate with its variation toeach data point in the original dataset. This effectively scaled the existing features (e.g.,order quantity, unit price) to reflect potential growth trends for 2024.
5. **Model Training and Prediction:** Finally, with the augmented dataset incorporating historical trends and projected variation, we trained the Random Forest model. This model was then used to predict future sales figures for 2024 based on the newly createddata.
   This indicates a positive trend with a projected sales growth from 2023 to2024. The percentage increase of 7.13% suggests a moderate but steady growth pattern.

Model	MSE	RMSE	MAE	MAPE	R2
Random Forest	59204457.659	7694.443	3770.628	0.015	0.713

SUGGESTIONS :

- The Company should explore alternatives to simple regression techniques for sales prediction, especially when dealing with short timeframes and limited historical data (a few years). Ensemble learning methods offer a promising approach.
- It must Implement ensemble learning techniques like Random Forest. These methods combine predictions from multiple models (e.g., decision trees) to achieve higher accuracy in sales forecasting.
- It should Move beyond internal sales data and explore the influence of external factors on sales. Integrate data like Indian CPI (Consumer Price Index) to understand inflation's impact on consumer spending.
- To Analyze industry-specific trends and competitor activity to predict market shifts that might affect sales.
- To Gauge public sentiment towards your brand and industry through social media analysis. This can provide insights into potential sales opportunities or risks.
- To Explore incorporating real-time sales data from point-of-sale systems or online platforms to create more dynamic sales forecasts that reflect current trends.

CONCLUSION :

This Study explored the potential of supervised machine learning algorithms to enhance sales forecasting accuracy at Sai Baba Auto Components. By analyzing various models, including Decision Tree Regression, Random Forest Regression, Gradient Boosting Machines, and Support Vector Regression, the project identified Random Forest Regression as the most effective approach for predicting sales based on historical data, achieving Lowest MSE, a RMSE, MAE, MAPE, and an Accuracy in R-squared value of 0.75. This project demonstrates the superiority of ensemble learning techniques like Random Forest Regression for sales forecasting at Sai Baba Auto Components, especially when dealing with historical data. By leveraging this model, the company can gain more accurate sales predictions, enabling better decision-making for resource allocation, sales planning, and achieving long-term sales success.

BIBLIOGRAPHY :

1. Nikhil et al. (2023). Sales Forecasting with Machine Learning: A Review.
2. Wu et al. (2022). Improving Sales Forecasting Accuracy with Machine Learning: A Comparative Analysis.
3. Abhishek et al. (2020). Machine Learning for Sales: Forecasting and Demand Planning.
4. Brodie et al. (2020). The Alignment Effect: How Machine Learning Can Drive Sustainable Advantage in Sales Forecasting.
5. Hamza et al. (2020). Forecasting Sales Using Machine Learning Techniques.
6. Pétteau, J-F. (2020). Time Series Forecasting with Python: Machine Learning, DeepLearning and Neural Networks.
7. Lessmann, S., et al. (2020). Explainable AI for Sales and Marketing: A Practitioner'sGuide.
8. Molnar, C. (2019). Interpretable Machine Learning.
9. Turban, E., et al. (2019). Business Analytics: Principles, Concepts, and Applications.
10. Mueller, J. P., & Massaron, L. (2019). Machine Learning for Dummies.
11. Géron, A. (2019). Feature Engineering for Machine Learning.
12. Deep Learning illustrated - Jon Krohn, Grant Beyleveld , Aglae Bassens 2020
13. Machine Learning for Absolute Beginners: A Plain English Introduction (Third Edition-Ethem Alpaydin, Addison-Wesley, 2014.
14. Hands-On Machine Learning with Scikit.-Learn, Keras & TensorFlow - Aurélien Géron, O'Reilly Media, 2017
15. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, WorkedExamples, and Case Studies - John D. Kelleher, Brian Mac Namee, and Aoife D'Arcy,Wiley, 2015.