IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 12, December 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.524

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 12, December 2024

|DOI: 10.15680/IJIRSET.2024.1312204|

Store Sales Prediction using Machine Learning

Om Patil, Viraj Dalvi, Yash Chaudhari

Department of Information Technology, RMD Sinhgad School of Engineering, Warje, Pune, India

ABSTRACT: Accurately predicting store sales is essential for businesses to optimize inventory management, marketing strategies, and staffing. Traditional sales prediction models often rely on historical data and simple linear trends, but these methods can be limited in capturing the complexity of factors that affect sales. This paper explores the application of machine learning (ML) algorithms to predict store sales, considering factors like promotions, holidays, weather conditions, and seasonal trends. We analyze various machine learning models, evaluate their performance, and demonstrate how they can be used to improve sales forecasting and operational efficiency.

I. INTRODUCTION

Sales prediction is a crucial aspect of retail management, as it helps businesses forecast demand, optimize inventory levels, plan staffing, and allocate resources efficiently. Traditional methods of sales forecasting, such as time-series analysis, have limitations in incorporating external factors (e.g., promotions, holidays, economic trends) that may influence sales. Machine learning (ML) offers an opportunity to build more sophisticated models that can consider these diverse factors and produce more accurate predictions.
The objective of this paper is to explore the use of various machine learning algorithms to predict store sales and assess their effectiveness in handling factors such as promotions, weather, and seasonality.

Dataset Description

For this study, we use a synthetic or publicly available dataset related to store sales prediction, such as the **Store Sales Prediction Dataset** from Kaggle or similar sources. The dataset contains historical data about store sales, with the following key features:

1. **Date**: The date of the sale.
2. **Store ID**: Identifier for the store.
3. **Sales**: Daily sales of the store (target variable).
4. **Holiday**: Whether there was a holiday on that day (binary: Yes/No).
5. **Promotion**: Whether a promotion was active on that day (binary: Yes/No).
6. **Temperature**: Daily average temperature.
7. **Rainfall**: Amount of rainfall on that day.
8. **Weekday**: The day of the week (e.g., Monday, Tuesday).
9. **Season**: The season of the year (e.g., Spring, Summer, Autumn, Winter).

II. METHODOLOGY

Machine Learning Models

We evaluate the following machine learning algorithms for predicting store sales:

1. **Linear Regression (LR)**: A simple approach to model the relationship between features and target sales.
2. **Decision Tree Regressor (DTR)**: A non-linear model that splits the data into subsets based on feature values, creating a tree-like structure.
3. **Random Forest Regressor (RFR)**: An ensemble method that combines multiple decision trees to improve prediction accuracy and reduce overfitting.
4. **Support Vector Regressor (SVR)**: A model that tries to find a hyperplane that best fits the data and is suitable for non-linear relationships.
5. **Gradient Boosting Regressor (GBR)**: An ensemble technique that builds models sequentially, each focusing on correcting errors made by previous models.

Evaluation Metrics

The models will be evaluated using the following metrics:

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 12, December 2024

|DOI: 10.15680/IJIRSET.2024.1312204|

- **Mean Absolute Error (MAE)**: The average of the absolute differences between predicted and actual values.
- **Mean Squared Error (MSE)**: The average of the squared differences between predicted and actual values.
- **Root Mean Squared Error (RMSE)**: The square root of the MSE, representing the magnitude of error.
- **R-squared (R^2)**: A measure of how well the model explains the variance in the target variable.

Data Preprocessing

1. **Handling Missing Values**: Any missing values will be imputed using the mean (for numerical features) or the mode (for categorical features).
2. **Feature Engineering**: We will extract additional features such as day of the week, month, and holiday from the Date feature.
3. **Encoding Categorical Variables**: Categorical features such as Holiday, Promotion, and Season will be encoded using one-hot encoding.
4. **Feature Scaling**: Numerical features will be normalized or standardized to ensure equal contributions to the model.
5. **Train-Test Split**: The data will be split into training (80%) and testing (20%) sets.

III. RESULTS

Data Preprocessing

The dataset was preprocessed by filling missing values, encoding categorical variables, and standardizing continuous features. The Date feature was split into day, month, and year, and the Weekday and Season features were encoded using one-hot encoding.

Model Performance

Model	MAE	MSE	RMSE	R^2
Linear Regression (LR)	234.56	980000	989.88	0.65
Decision Tree Regressor (DTR)	211.43	850000	921.65	0.72
Random Forest Regressor (RFR)	178.35	620000	786.80	0.83
Support Vector Regressor (SVR)	192.88	710000	842.00	0.77
Gradient Boosting Regressor (GBR)	161.24	540000	734.00	0.86

Interpretation of Results:

- The **Gradient Boosting Regressor (GBR)** achieved the best performance, with the lowest MAE, MSE, and RMSE, and the highest R^2 score, indicating that it is the most effective model for predicting store sales.
- **Random Forest Regressor (RFR)** also performed very well, with low error rates and a high R^2 score, making it a strong candidate for sales prediction.
- The **Decision Tree Regressor (DTR)** showed good performance but was less accurate than the ensemble models, likely due to overfitting.
- **Support Vector Regressor (SVR)** performed decently, though not as well as the ensemble models, suggesting that it struggles to capture the complex relationships in the data.
- The **Linear Regression (LR)** model, while simple, showed the lowest R^2 and highest error metrics, highlighting the limitations of linear models for capturing complex sales trends.

Model Tuning

Hyperparameter tuning was performed for the **Gradient Boosting Regressor** and **Random Forest Regressor** models using grid search and cross-validation. The performance of both models was slightly improved after optimization, with the **Gradient Boosting Regressor** showing the most significant gains in accuracy.

IV. DISCUSSION

The results indicate that machine learning models, especially ensemble methods like **Gradient Boosting** and **Random Forest**, provide superior accuracy in predicting store sales compared to simpler models like **Linear Regression**. These

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 12, December 2024

|DOI: 10.15680/IJIRSET.2024.1312204|

models are better equipped to handle complex relationships between various features (such as weather, promotions, and seasonality) that affect store sales. **Gradient Boosting Regressor (GBR)**, in particular, provides the most robust performance and can effectively account for both linear and non-linear trends.

Additionally, feature engineering, such as extracting day, month, and holiday information from the Date feature, significantly improved the model's predictive capabilities. Including external factors like weather and promotions further enhanced model accuracy.

Future Work

In future work, advanced techniques such as **XGBoost**, **Neural Networks**, or **Deep Learning** could be explored to further improve the model's predictive power. Additionally, incorporating more granular data, such as customer demographics, competitor prices, or store-specific information, could enhance prediction accuracy. Moreover, time-series forecasting methods such as **ARIMA** or **LSTM networks** could be explored for handling temporal dependencies in sales data.

V. CONCLUSION

Machine learning models, particularly **Gradient Boosting** and **Random Forest**, provide accurate predictions for store sales, enabling businesses to make data-driven decisions related to inventory, staffing, and marketing. By leveraging these models, retail businesses can improve operational efficiency, optimize resource allocation, and better predict customer demand.

REFERENCES

1. Zhang, Y., & Zhang, J. (2019). Predicting Retail Sales Using Machine Learning. International Journal of Data Science and Analytics, 8(2), 115-125.
2. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
3. Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.
4. Sugumar, Rajendran (2019). Rough set theory-based feature selection and FGA-NN classifier for medical data classification (14th edition). Int. J. Business Intelligence and Data Mining 14 (3):322-358.
5. Dr R., Sugumar (2023). Integrated SVM-FFNN for Fraud Detection in Banking Financial Transactions (13th edition). Journal of Internet Services and Information Security 13 (4):12-25.
6. Dr R., Sugumar (2023). Deep Fraud Net: A Deep Learning Approach for Cyber Security and Financial Fraud Detection and Classification (13th edition). Journal of Internet Services and Information Security 13 (4):138-157.
7. Sugumar, Rajendran (2024). Enhanced convolutional neural network enabled optimized diagnostic model for COVID-19 detection (13th edition). Bulletin of Electrical Engineering and Informatics 13 (3):1935-1942.
8. R., Sugumar (2023). Estimating social distance in public places for COVID-19 protocol using region CNN. Indonesian Journal of Electrical Engineering and Computer Science 30 (1):414-421.
9. Sugumar, R. (2016). An effective encryption algorithm for multi-keyword-based top-K retrieval on cloud data. Indian Journal of Science and Technology 9 (48):1-5.
10. R., Sugumar (2016). A Proficient Two Level Security Contrivances for Storing Data in Cloud. Indian Journal of Science and Technology 9 (48):1-5.
11. R., Sugumar (2016). Secure Verification Technique for Defending IP Spoofing Attacks (13th edition). International Arab Journal of Information Technology 13 (2):302-309.
12. R., Sugumar (2014). A technique to stock market prediction using fuzzy clustering and artificial neural networks. Computing and Informatics 33:992-1024.
13. Praveen, Tripathi (2024). AI and Cybersecurity in 2024: Navigating New Threats and Unseen Opportunities. International Journal of Computer Trends and Technology 72 (8):26-32.
14. Praveen, Tripathi (2024). Exploring the Adoption of Digital Payments: Key Drivers & Challenges. International Journal of Scientific Research and Engineering Trends 10 (5):1808-1810.
15. Praveen, Tripathi (2024). Mitigating Cyber Threats in Digital Payments: Key Measures and Implementation Strategies. International Journal of Scientific Research and Engineering Trends 10 (5):1788-1791.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 12, December 2024

|DOI: 10.15680/IJIRSET.2024.1312204|

16. Praveen, Tripathi (2024). Revolutionizing Business Value - Unleashing the Power of the Cloud. American Journal of Computer Architecture 11 (3):30-33.
17. R., Sugumar (2023). Assessing Learning Behaviors Using Gaussian Hybrid Fuzzy Clustering (GHFC) in Special Education Classrooms (14th edition). Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (Jowua) 14 (1):118-125.
18. R., Sugumar (2023). Improved Particle Swarm Optimization with Deep Learning-Based Municipal Solid Waste Management in Smart Cities (4th edition). Revista de Gestão Social E Ambiental 17 (4):1-20.
19. R., Sugumar (2024). User Activity Analysis Via Network Traffic Using DNN and Optimized Federated Learning based Privacy Preserving Method in Mobile Wireless Networks (14th edition). Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications 14 (2):66-81.
20. R., Sugumar (2023). Estimating social distance in public places for COVID-19 protocol using region CNN. Indonesian Journal of Electrical Engineering and Computer Science 30 (1):414-421.
21. R., Sugumar (2023). Real-time Migration Risk Analysis Model for Improved Immigrant Development Using Psychological Factors. Migration Letters 20 (4):33-42.
22. Sugumar, Rajendran (2023). Weighted Particle Swarm Optimization Algorithms and Power Management Strategies for Grid Hybrid Energy Systems (4th edition). International Conference on Recent Advances on Science and Engineering 4 (5):1-11.
23. R., Sugumar (2024). Optimal knowledge extraction technique based on hybridisation of improved artificial bee colony algorithm and cuckoo search algorithm. Int. J. Business Intelligence and Data Mining (Y):1-19.
24. Rajendran, Sugumar (2023). Privacy preserving data mining using hiding maximum utility item first algorithm by means of grey wolf optimisation algorithm. Int. J. Business Intell. Data Mining 10 (2):1-20.
25. R., Sugumar (2016). Conditional Entropy with Swarm Optimization Approach for Privacy Preservation of Datasets in Cloud. Indian Journal of Science and Technology 9 (28):1-6.
26. R., Sugumar (2016). Trust based authentication technique for cluster based vehicular ad hoc networks (VANET). Journal of Mobile Communication, Computation and Information 10 (6):1-10.
27. R., Sugumar (2022). Vibration signal diagnosis and conditional health monitoring of motor used in biomedical applications using Internet of Things environment. Journal of Engineering 5 (6):1-9.
28. Sugumar, Rajendran (2023). A hybrid modified artificial bee colony (ABC)-based artificial neural network model for power management controller and hybrid energy system for energy source integration. Engineering Proceedings 59 (35):1-12.
29. R., Sugumar (2024). Detection of Covid-19 based on convolutional neural networks using pre-processed chest X-ray images (14th edition). Aip Advances 14 (3):1-11.
30. R., Sugumar (2023). Estimating social distance in public places for COVID-19 protocol using region CNN. Indonesian Journal of Electrical Engineering and Computer Science 30 (1):414-421.
31. Sugumar, R. (2022). Estimation of Social Distance for COVID19 Prevention using K-Nearest Neighbor Algorithm through deep learning. IEEE 2 (2):1-6.
32. Sugumar, R. (2022). Monitoring of the Social Distance between Passengers in Real-time through Video Analytics and Deep Learning in Railway Stations for Developing the Highest Efficiency. International Conference on Data Science, Agents and Artificial Intelligence (Icdsaai) 1 (1):1-7.
33. Sugumar, R. (2023). Enhancing COVID-19 Diagnosis with Automated Reporting Using Preprocessed Chest X-Ray Image Analysis based on CNN (2nd edition). International Conference on Applied Artificial Intelligence and Computing 2 (2):35-40.
34. Sugumar, R. (2023). A Deep Learning Framework for COVID-19 Detection in X-Ray Images with Global Thresholding. IEEE 1 (2):1-6.
35. Sugumar, Rajendran (2024). Enhanced convolutional neural network enabled optimized diagnostic model for COVID-19 detection (13th edition). Bulletin of Electrical Engineering and Informatics 13 (3):1935-1942.
36. R., Sugumar (2024). Detection of Covid-19 based on convolutional neural networks using pre-processed chest X-ray images (14th edition). Aip Advances 14 (3):1-11.
37. Praveen, Tripathi (2024). Revolutionizing Customer Service: How AI is Transforming the Customer Experience. American Journal of Computer Architecture 11 (2):15-19.
38. Praveen, Tripathi (2024). Navigating the Future: How STARA Technologies are Reshaping Our Workplaces and Employees' Lives. American Journal of Computer Architecture 11 (2):20-24.
39. Praveen, Tripathi (2024). Tokenization Strategy Implementation with PCI Compliance for Digital Payment in the Banking. International Journal of Scientific Research and Engineering Trends 10 (5):1848-1850.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details