International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

FEATURE ENGINEERING FOR SALES DATASET

[1]J.Priyanka,[2]P.Meghana,[3]CH.Shravani,[4]K.Kamaleshwar reddy

[1]Assistant Professor, [234]Students

Department of CSE(Data Science)

Siddhartha institute of technology & sciences,narapally

jalagapriyanka.cse@siddhartha.co.in, 23TQ1A6712@siddhartha.co.in, 23TQ1A6737@siddhartha.co.in, 23TQ1A6747@siddhartha.co.in

ABSTRACT

In today's data-driven business environment, organizations generate a large volume of sales data from various sources such as retail transactions, e-commerce platforms, and customer purchases. However, raw sales data often contains missing values, inconsistent formats, and irrelevant attributes, which makes it difficult to analyse and derive meaningful insights. Feature engineering plays a crucial role in transforming raw data into meaningful and useful features that improve data analysis and decision making processes. The main objective of this project is to perform feature engineering on a sales dataset to enhance the quality of data and extract valuable business insights. The project involves several stages including data collection, data preprocessing, feature creation, exploratory data analysis (EDA), and data visualization. Various data preprocessing techniques such as handling missing values, removing duplicate records, and transforming data into appropriate formats are applied. Python programming language is used for performing feature engineering and data analysis with the help of libraries such as Pandas, NumPy, Matplotlib, and Seaborn. New features such as revenue, profit margin, and category-based analysis are generated from the raw dataset. Exploratory data analysis is conducted to understand sales patterns, product performance, and revenue distribution. Furthermore, Tableau is used to develop interactive dashboards that visually represent sales trends, profitability analysis, and category distribution. The results obtained from this project help businesses understand sales performance, identify profitable products, and support data-driven decision making. This project demonstrates the importance of feature engineering in transforming raw sales data into meaningful insights for business analytics

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

I INTRODUCTION

In the modern digital era, businesses generate a huge amount of data through sales transactions, online purchases, customer interactions, and product records. This data contains valuable information that can help organizations understand customer behavior, improve product performance, and increase overall business profitability. However, raw data collected from different sources is often unstructured and contains inconsistencies such as missing values, duplicate records, and irrelevant attributes. Therefore, it becomes necessary to process and transform this raw data into a structured format before performing analysis. Feature Engineering is one of the most important steps in data analysis and machine learning. It involves transforming raw data into meaningful features that help improve the performance of analytical models and provide deeper insights. Proper feature engineering helps in identifying hidden patterns, relationships, and trends in the dataset. In this project, a Sales Dataset is analysed to understand sales performance and generate useful business insights. The dataset contains information such as product details, quantity sold, revenue, and profit values. Through the application of feature engineering techniques, new variables are created from the existing dataset to improve data understanding and analysis. Python programming language is used for performing data preprocessing and feature engineering with the help of powerful libraries such as Pandas, NumPy, Matplotlib, and Seaborn. These libraries help in handling large datasets efficiently and generating various visualizations to understand the data better. After processing the dataset, Exploratory Data Analysis (EDA) is performed to analyze sales patterns, revenue trends, and product category performance. Various graphs and charts are created to visualize the relationships between different variables. Finally, the processed dataset is visualized using Tableau, which provides an interactive dashboard for analysing sales performance. The dashboard helps businesses identify high-performing products, profitable categories, and seasonal sales trends. This project demonstrates how feature engineering and data visualization can transform raw sales data into meaningful insights that support better business decision-making.

II LITERATURE SURVEY

Sales data analysis has become an important area of research in the field of data analytics and business intelligence. Organizations generate large volumes of sales data through retail transactions, online purchases, and customer interactions. Analyzing this data helps businesses understand customer behavior, identify profitable products, and improve decision-making processes. However, raw sales datasets often contain missing values, redundant information, and inconsistent formats. Researchers have emphasized the importance of data

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

preprocessing, feature engineering, exploratory data analysis (EDA), and visualization tools for extracting meaningful insights from sales data.

1. A Study on Data Preprocessing and Feature Engineering in Data

Several research studies highlight the importance of data preprocessing and feature engineering as essential steps in data analysis. Raw datasets collected from real-world environments often contain missing values, duplicate records, and inconsistent data formats. According to studies in data science, preprocessing techniques such as data cleaning, normalization, transformation, and feature creation help improve the quality of datasets and make them suitable for analysis.

2. A Study on Exploratory Data Analysis and Visualization

Exploratory Data Analysis (EDA) plays a crucial role in understanding the structure and patterns present in a dataset. Researchers emphasize that visualization techniques help analysts identify hidden relationships, correlations, and trends within large datasets. Tools such as Matplotlib, Seaborn, and Plotly in Python are widely used to generate visual representations like bar charts, line graphs,

histograms, and scatter plots.

III SYSTEM ANALYSIS

Feature engineering for a sales dataset is a critical step in system analysis, as it transforms raw transactional data into meaningful inputs that improve model performance and business insights. In this process, existing variables such as date, product ID, customer details, and sales amount are used to create new informative features like monthly sales trends, seasonal indicators, customer purchase frequency, average order value, and product demand patterns. Time-based features (day, week, quarter), lag features (previous sales), and rolling averages help capture trends and seasonality, which are essential for forecasting. Additionally, categorical variables like region or product category are encoded into numerical formats to make them usable for machine learning models. Proper handling of missing values, outliers, and normalization ensures data consistency and reliability. Overall, feature engineering enhances the analytical capability of the system by enabling more accurate predictions, better customer segmentation, and improved decision-making in sales strategy and inventory management.

Existing system

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

The existing system for handling sales datasets is mostly manual or relies on basic data processing techniques with limited analytical capabilities. Sales data is typically stored in spreadsheets or simple databases, where analysis is performed using basic tools like reports and charts. This system lacks advanced feature engineering, meaning it does not fully utilize important patterns such as seasonality, customer behavior, or product trends. As a result, decision-making is often based on historical summaries rather than predictive insights. Additionally, the system may suffer from issues like inconsistent data formats, missing values, and lack of automation, leading to inefficiencies and errors. Overall, the existing system is time-consuming, less accurate, and not capable of supporting intelligent forecasting or data-driven business strategies.

Disadvantages of existing system

- Manual data handling is time-consuming and prone to human errors
- Lack of automation in data processing and analysis
- No advanced feature engineering to capture hidden patterns
- Poor handling of missing or inconsistent data
- Limited ability to analyze customer behavior and trends

Proposed system

The proposed system introduces an advanced and automated approach for handling sales datasets by incorporating feature engineering and machine learning techniques. It transforms raw sales data into meaningful features such as time-based trends, customer purchasing behavior, and product demand patterns, enabling deeper analysis. The system ensures proper data preprocessing, including handling missing values, removing outliers, and normalizing data for consistency. With the integration of predictive models, it can forecast future sales, identify high-performing products, and support data-driven decision-making. Additionally, the system improves efficiency through automation, reduces manual effort, and provides real-time insights through dashboards and reports. Overall, the proposed system enhances accuracy, scalability, and business intelligence, making it more effective for modern sales analysis and strategic planning.

Advantages of proposed system

- Automated data processing reduces manual effort and saves time
- Improved accuracy through proper data preprocessing and feature engineering

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

- Ability to identify hidden patterns and trends in sales data
- Supports predictive analysis and sales forecasting
- Better understanding of customer behavior and preferences

IV METHODOLOGY

The methodology of the Feature Engineering for Sales Dataset project explains the systematic process followed to transform raw sales data into meaningful insights. The project follows a structured workflow that includes data collection, data preprocessing, feature engineering, exploratory data analysis, visualization, and interpretation of results. Each step plays an important role in ensuring accurate analysis and meaningful insights. The methodology of this project involves several steps to prepare and improve the sales dataset using feature engineering techniques. First, the sales dataset is collected from a reliable source and loaded into the analysis environment. After collecting the data, the dataset is carefully examined to understand its structure, variables, and possible issues such as missing values or incorrect entries. The next step is data preprocessing, where the dataset is cleaned by removing duplicate records, handling missing values, and correcting inconsistent data. This step ensures that the dataset becomes accurate and reliable for further analysis. After cleaning the data, exploratory data analysis is performed to understand the distribution of variables and identify patterns or relationships in the sales data. Once the data is properly understood, feature engineering techniques are applied. In this step, new useful features are created from existing variables, such as calculating total sales, sales growth, or extracting time-related features like month or year from the date column. These newly created features help improve the quality of the dataset and make it more informative for analysis. After creating the features, feature selection is performed to choose the most important variables that contribute to better analysis and prediction. Unnecessary or less useful features are removed to reduce complexity and improve model performance. Finally, the processed dataset is prepared for further analysis, visualization, or machine learning models. The results obtained from the improved dataset help in identifying sales trends, understanding business performance, and supporting better decision-making.

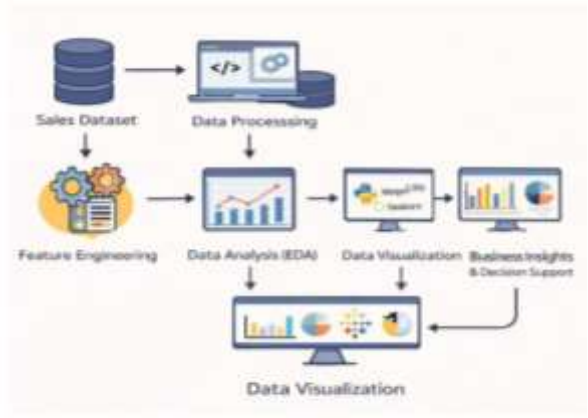
International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

System Architecture



The system architecture for a sales dataset with feature engineering is designed to efficiently handle data collection, processing, analysis, and prediction in a structured manner. It consists of multiple layers that work together to transform raw data into meaningful insights. The first layer is the data collection layer, where sales data is gathered from various sources such as databases, APIs, or transaction systems. The next layer is the data preprocessing layer, which handles cleaning tasks like removing duplicates, managing missing values, and standardizing formats.

Following this, the feature engineering layer plays a key role by creating new meaningful features such as time-based attributes, customer purchase frequency, and product performance metrics. These features are then passed to the model layer, where machine learning algorithms are applied for tasks like sales prediction and trend analysis. The system also includes a storage layer to save processed data and model outputs efficiently.

Finally, the visualization and reporting layer presents insights through dashboards, charts, and reports, helping stakeholders make informed decisions. Overall, this architecture ensures a smooth flow of data, improves accuracy, enables automation, and supports scalable and intelligent sales analysis.

V RESULTS&OUTPUT

The results of this project demonstrate how feature engineering and data visualization techniques help transform raw sales data into meaningful insights. After preprocessing and feature engineering, the dataset was analysed using Python libraries such as Pandas, Matplotlib, and Seaborn. Various visualizations were generated to understand sales patterns, revenue trends, and product category performance. Graphs such as bar

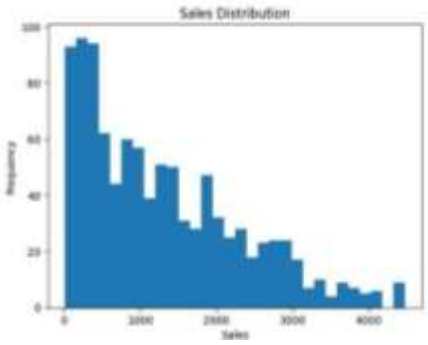
International Journal of

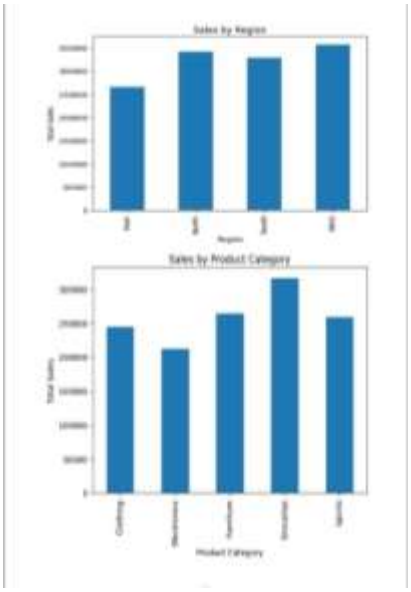
DATA SCIENCE AND IOT MANAGEMENT SYSTEM

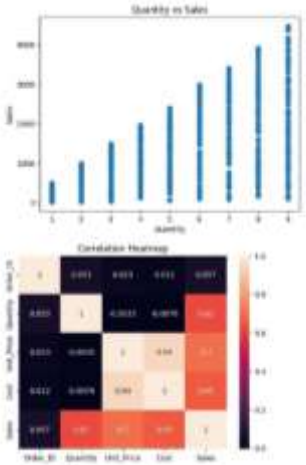
Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

charts, line charts, scatter plots, and pie charts were created to analyse relationships between sales variables. The analysis helped identify important business insights such as monthly sales trends, profitable product categories, and revenue distribution across different products. These insights help organizations understand sales performance and make better business decisions







VI CONCLUSION

This project focuses on performing feature engineering and exploratory data analysis on a sales dataset to extract meaningful insights that support business decision-making. Raw sales data often contains inconsistencies and lacks structure, which makes it difficult to analyze directly. Through data preprocessing and feature engineering techniques, the dataset was transformed into a structured format suitable

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

for analysis. Using Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn, the dataset was analyzed to identify important sales trends and relationships between different variables. Visualizations were generated to understand revenue patterns, product performance, and sales distribution. Additionally, an interactive dashboard was created using Tableau to present the results in a clear and user-friendly format. The dashboard enables users to easily explore the sales dataset and identify key business insights. Overall, the project demonstrates how feature engineering and visualization techniques can convert raw sales data into valuable information that helps organizations make data-driven decisions and improve business performance. Furthermore, this project highlights the importance of combining data preprocessing, feature engineering, and visualization techniques to extract valuable information from raw datasets. By applying systematic analysis and creating interactive dashboards, the project demonstrates how organizations can easily interpret complex sales data and identify important business trends. The insights obtained from the analysis can assist companies in improving product strategies, optimizing resource allocation, and enhancing overall business performance.

REFERENCE

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, “Real-Time Object Detection in Drone Surveillance Using YOLOv5,” in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, “Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks,” in

Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.