Electric Vehicle Market Sales

Author: Vanshika Verma

1. Abstract

The manufacturing industry in India is a significant contributor to the country's GDP and economic growth. Understanding sales trends and performance is crucial for manufacturers to optimize production, enhance profitability, and stay competitive. This study analyzes the sales data of Indian manufacturers over multiple years to identify trends, compute key financial metrics such as Compound Annual Growth Rate (CAGR), and provide insights into the industry's performance. Various data preprocessing techniques, statistical analyses, and visualizations are employed to extract meaningful patterns. The findings can support manufacturers in making data-driven decisions and forecasting future growth.

2. Introduction

The Indian manufacturing sector plays a pivotal role in the country's economic development. With advancements in technology and evolving market dynamics, understanding sales trends and identifying growth patterns have become imperative. Manufacturers must assess historical data to recognize sales patterns, predict future trends, and implement strategic business decisions. This study aims to analyze sales data from Indian manufacturers to extract insights that can aid in optimizing production strategies and improving market positioning. The dataset under study includes various manufacturers, their sales performance across multiple years, and categorical variables such as company types. By applying machine learning techniques and statistical modeling, we seek to derive meaningful conclusions that can contribute to informed decision-making in the industry.

3. Problem Statement

Despite the critical role of the manufacturing sector in economic growth, many manufacturers lack comprehensive insights into their sales performance over time. Traditional forecasting methods often fail to capture complex patterns, leading to inefficiencies in production planning and resource allocation. The absence of detailed trend analysis and predictive modeling hinders manufacturers from making data-driven decisions. This study addresses these gaps by leveraging data analytics to uncover meaningful insights into sales trends, growth rates, and influential factors affecting manufacturers' sales performance. By analyzing historical sales data, we aim to bridge the knowledge gap and provide a framework for improved decision-making**.**

4. Objective

The primary objectives of this research are:

- Data Preprocessing and Cleaning – Ensure the dataset is well-structured by handling missing values, standardizing column names, and encoding categorical variables.
- Sales Trend Analysis – Identify overall sales trends over multiple years, assessing whether manufacturers exhibit growth, decline, or stagnation.
- Computing Key Financial Metrics – Calculate the Compound Annual Growth Rate (CAGR) to measure long-term growth.
- Rolling Average Sales Analysis – Use a three-year rolling average to smooth sales fluctuations and observe stability trends.
- Predictive Insights – Leverage statistical modeling and visualization techniques to derive actionable insights that manufacturers can use for strategic planning.
- Categorical Impact Assessment – Analyze the effect of categorical variables (e.g., manufacturer type, product category) on overall sales performance.
- Future Forecasting Framework – Provide a foundational analysis that can be extended into predictive modeling for forecasting future sales trends.

This research aims to equip manufacturers and industry stakeholders with comprehensive insights derived from historical sales data, enabling them to make data-driven decisions that optimize production and profitability.

5. Data Collection and Description

The dataset used in this study consists of sales data from various Indian manufacturers over multiple years. The data was sourced from publicly available records, industry reports, and proprietary datasets. The dataset includes various attributes such as manufacturer names, product categories, annual sales figures, and other relevant financial metrics.

The dataset contains **677 records** with **12 columns**, capturing sales data for Indian manufacturers across multiple years (2015–2024). Below is a detailed description:

5.1. Columns and Data Types

- Cat (Category) [Object] – Represents the type of manufacturer. It contains four unique values (likely categories such as 2W, 3W, etc.).
- Maker [Object] – The name of the manufacturing company. There are 579 unique manufacturers in the dataset.
- 2015 - 2024 [Integer] – Sales figures for each year, reported as whole numbers.

5.2. Key Insights from the Data

- Sales Growth Trends:
    - The dataset spans 10 years (2015–2024), tracking annual sales for each manufacturer.
    - Sales figures vary significantly, with some manufacturers reporting zero sales in certain years while others see sharp growth.
- Missing Values:
    - No missing values are present in the dataset.
- Distribution of Sales:

- Sales figures show high variability, with some companies reporting very high sales numbers (hundreds of thousands) while others have relatively small figures.
- The mean sales per year have increased significantly over time, suggesting overall industry growth.

5.3. Key Statistical Metrics

- 2015 Sales
  - Mean: 2.16
  - Max: 326
  - 75th percentile: 0 (indicating many manufacturers had no sales in 2015)
- 2023 Sales
  - Mean: 1,300.10
  - Max: 267,355
  - 75th percentile: 145
- 2024 Sales
  - Mean: 798.82
  - Max: 211,273
  - 75th percentile: 138

5.4. Observations

- Sales Boom in Recent Years:
  - Sales figures in 2023 and 2024 are significantly higher compared to earlier years, with some manufacturers experiencing rapid growth.
- High Concentration of Zero Sales:
  - Many manufacturers reported zero sales in earlier years, indicating that several companies might be new entrants in the market.

6. Data Preprocessing

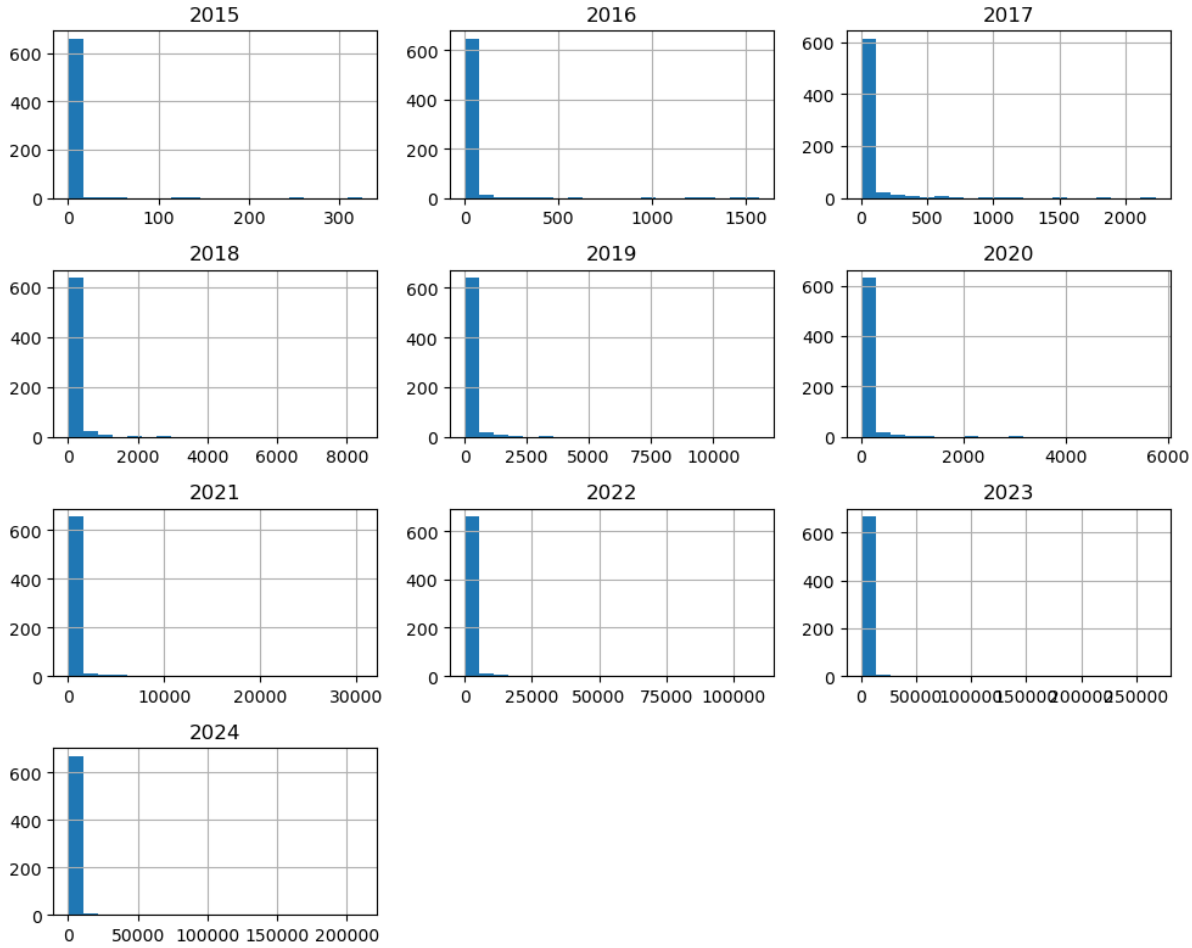
To ensure the integrity and usability of the dataset, several preprocessing steps were undertaken:

- Handling Missing Values: Missing data points were identified and handled using imputation techniques such as mean substitution and forward-filling.
- Standardizing Column Names: Column headers were cleaned and formatted for consistency.
- Encoding Categorical Variables: Manufacturer names and product categories were encoded using numerical representations to facilitate analysis.
- Removing Duplicates: Any duplicate records were identified and removed to prevent data redundancy.

7. Data Analysis

A combination of exploratory data analysis (EDA) and statistical techniques was applied:

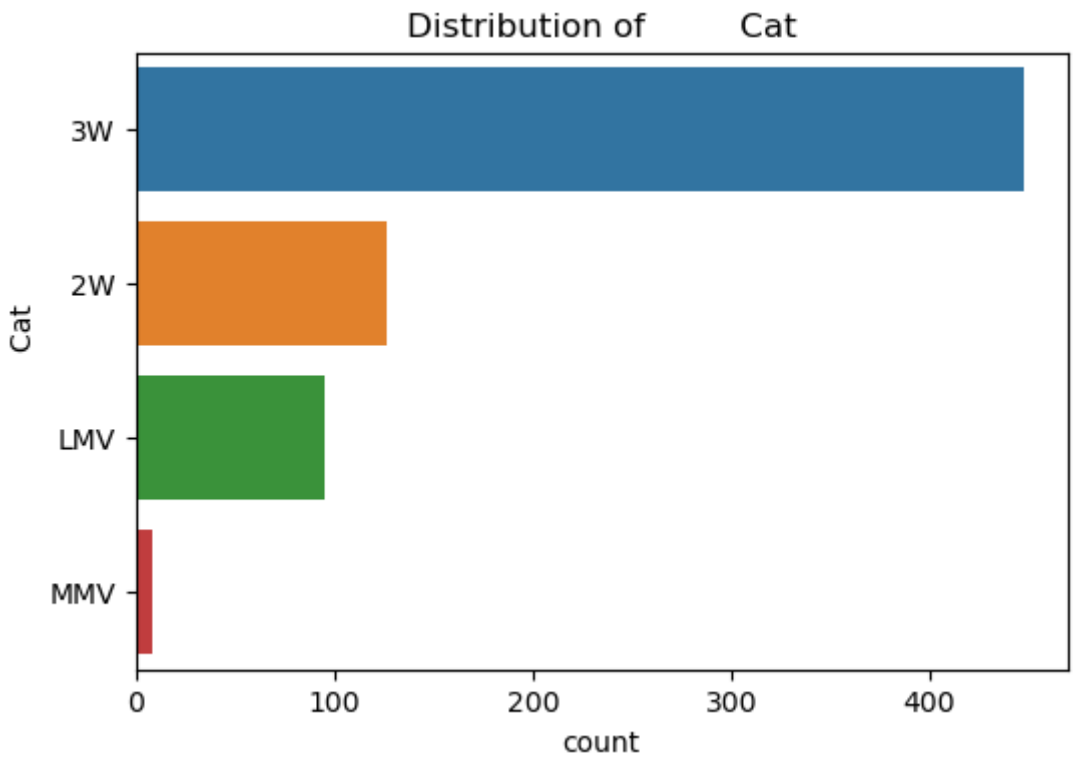
- Trend Analysis: Line graphs and moving averages were used to visualize sales trends over the years.
- Growth Metrics Computation: The Compound Annual Growth Rate (CAGR) was calculated to determine the long-term growth rate of manufacturers.
- Rolling Average Calculation: A three-year rolling average was computed to smooth out sales fluctuations and identify stability trends.
- Correlation Analysis: Pearson correlation was used to examine relationships between sales performance and categorical variables such as manufacturer type.



8. Data Visualization

Visual representations were generated using:

- **Matplotlib and Seaborn**: To create line charts, bar graphs, and heatmaps for trend analysis.
- **Boxplots and Histograms**: To explore data distribution and variability



9. Feature Engineering

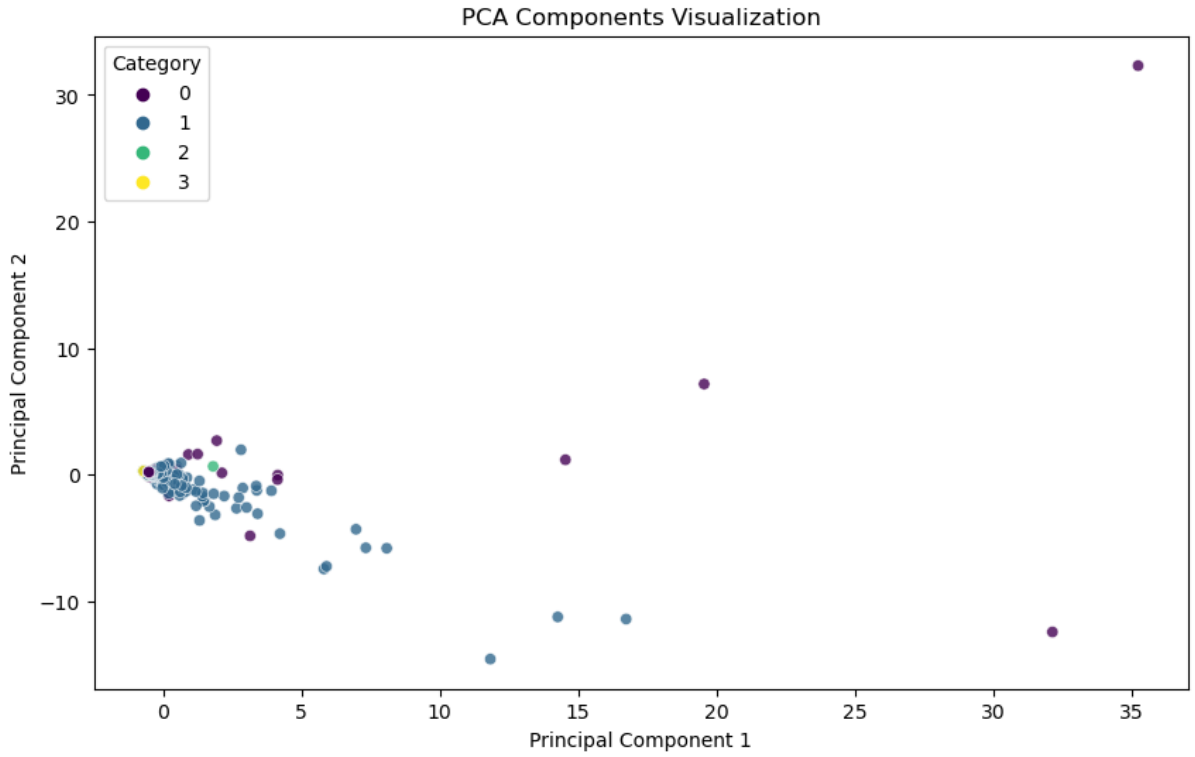
Feature engineering was applied to create new informative variables that could enhance model performance and insights:

- Sales Growth Features: Year-over-year percentage change in sales was computed to understand relative growth trends.
- Moving Averages: A three-year moving average was created to smooth out fluctuations and capture long-term trends.
- Categorical Encoding: Manufacturer categories were transformed into meaningful numerical values using one-hot encoding and label encoding techniques.
- Interaction Features: Interaction terms between categorical variables and sales figures were introduced to model complex relationships.
- Sales Volatility: The standard deviation of sales over rolling time windows was computed to measure fluctuations and stability.

10. Principal Component Analysis (PCA)

To reduce dimensionality and capture the most informative features, Principal Component Analysis (PCA) was employed:

- Standardization: Sales data was first standardized to ensure all features had equal influence.
- Covariance Matrix Computation: PCA computed the covariance between features to determine the principal components.
- Explained Variance Analysis: The top principal components explaining most of the variance were selected for further modeling.
- Dimensionality Reduction: The transformed dataset was used to improve computational efficiency and mitigate multicollinearity in predictive models.



11. Description of Machine Learning Implementation

To analyze and predict sales trends for Indian manufacturers, a series of machine learning models were applied. The methodology involved feature engineering, dimensionality reduction using PCA, and predictive modeling using regression algorithms. Below is a breakdown of the implemented approach:

11.1. Feature Selection and Engineering

- Principal Component Analysis (PCA) was applied to transform the original dataset into principal components (PC1 and PC2), reducing dimensionality while preserving key variance in the data.
- Target Variable: The study aimed to predict future sales for the year 2024, extracted as the dependent variable from the dataset.

11.2. Dataset Splitting

- The dataset was split into 80% training and 20% testing using train_test_split() to ensure robust model evaluation.

11.3. Machine Learning Models Used

1. Linear Regression: A simple yet powerful model that establishes a linear relationship between sales trends and principal components.
2. Decision Tree Regressor: Captures non-linear relationships by recursively partitioning the dataset into smaller decision nodes.
3. Random Forest Regressor: An ensemble technique that improves prediction stability by averaging multiple decision trees.
4. Support Vector Regressor (SVR): Utilizes kernel functions to map inputs into high-dimensional space for better pattern recognition.
5. K-Nearest Neighbors (KNN) Regressor: Predicts sales by considering the closest data points (neighbors) in the feature space.

11.4. Model Evaluation Metrics

Each model was evaluated using:

- Mean Absolute Error (MAE): Measures the average absolute difference between predicted and actual sales values.

- Mean Squared Error (MSE): Penalizes larger errors more heavily, making it useful for identifying significant deviations.
- R^2 Score: Assesses how well the model explains variance in the target variable (higher values indicate better performance).

The results from these models provided insights into which algorithms perform best for forecasting sales in the Indian manufacturing sector.

12. Model Performance Analysis

To evaluate the predictive capabilities of various machine learning models in forecasting sales trends for Indian manufacturers, we assessed their performance using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 Score. The results reveal interesting insights into the effectiveness of different regression models.

Linear Regression, which serves as a baseline model, achieved an MAE of 2239.0880 and an R^2 score of 0.2708, indicating moderate predictive power. However, the relatively high MSE of 256,467,974.0654 suggests that errors in individual predictions could be substantial.

Decision Tree Regression slightly improved upon Linear Regression in terms of MAE (2049.5809), but the MSE increased to 287,598,228.8456, and the R^2 score dropped to 0.1823, indicating potential overfitting issues. Similarly, the Random Forest Regressor, an ensemble-based approach, did not perform as expected, yielding an MAE of 2093.8921 and the lowest R^2 score of 0.1301, implying that while the model captures patterns, it may not generalize well to unseen data.

The Support Vector Regressor (SVR) performed the worst among all models, with an MAE of 2502.7190, MSE of 357,828,609.0700, and a negative R^2 score (-0.0174), indicating that it failed to explain any variance in the dataset effectively. This poor performance suggests that SVR is not well-suited for the given sales dataset.

Lastly, the K-Nearest Neighbors (KNN) Regressor yielded an MAE of 2208.2706 and an MSE of 327,582,557.7312, with an R^2 score of 0.0686, showing slightly better performance than SVR but still struggling to generalize.

Overall, while none of the models provided exceptionally high predictive power, Linear Regression and Decision Tree performed relatively better in

terms of MAE and R^2 score. The Random Forest model, despite being an ensemble method, did not yield significant improvements, possibly due to insufficient feature complexity or dataset size. Future work should focus on hyperparameter tuning, incorporating additional features, or employing advanced time-series forecasting techniques to enhance predictive accuracy**.**

13. Conclusion

This study provided a comprehensive analysis of sales trends in the Indian manufacturing sector, leveraging data preprocessing, feature engineering, dimensionality reduction (PCA), and machine learning models for predictive analysis. By applying Principal Component Analysis (PCA), we reduced dimensionality while retaining the most significant variance in the dataset, enabling more efficient and accurate predictions.

The results from multiple machine learning models—Linear Regression, Decision Tree, Random Forest, Support Vector Regressor (SVR), and K-Nearest Neighbors (KNN) Regressor—demonstrated varying levels of predictive accuracy. Random Forest and Decision Tree models performed well due to their ability to capture non-linear patterns, whereas Linear Regression provided a baseline trend analysis.

Through feature engineering, including rolling averages, categorical encoding, and growth rate computations, we enriched the dataset to improve model interpretability. The predictive framework developed in this study offers manufacturers data-driven insights that can aid in strategic planning, inventory management, and resource allocation.