SHRI RAMDEOBABA COLLEGE OF ENGINEERING AND MANAGEMENT, NAGPUR.

Machine Learning

TEACHER'S ASSESSMENT

(7TH SEM-A, SESSION 2024-2025, ECT-453-3)

"Sales Prediction"

Submitted By

Khushi Agrawal (Roll No.-32)
Nishad Gharote (Roll No.-44)
Piyush Pandit (Roll No.-45)
Shruti Desuwar (Roll No.-63)

Department of Electronics and Communication Engineering

1. Introduction:

The increasing competition in the market needs accurate predictions of sales to formulate effective marketing strategies. This project aims to develop a predictive model that predicts sales based on expenditures on various marketing platforms, namely TV, radio, and newspapers.

The primary objectives are:

- To define the relationship between marketing spend and sales.
- To implement predictive models using various Machine Learning Algorithms.
- To evaluate the models' performance and draw actionable insights.

Problem Statement: Despite substantial investments in marketing, many companies frequently find it challenging to determine how effective these expenditures are in driving sales. This study addresses this gap by analyzing historical data to predict future sales based on advertising spend.

2. Existing Approaches or algorithm:

Several methods exist for sales prediction, ranging from statistical techniques to machine learning algorithms.

- **Simple Linear Regression**: This model predicts sales based on a single feature (e.g., the amount spent on marketing). The relationship is represented by a straight line, where the slope indicates how much sales are expected to change with each unit increase in the feature.
- **Multiple Linear Regression:** This model extends simple linear regression to include multiple features (e.g., marketing spend, economic indicators, product price). It captures the combined effect of several factors on sales.
- **Machine Learning Techniques:** More complex algorithms like decision trees and neural networks can yield higher accuracy but require extensive tuning.
- **Causal Models:** Causal models are used to figure out the actual causes behind changes in sales. Unlike methods that just identify patterns or trends, causal models aim to understand which factors directly influence sales by using economic theory and statistical methods to estimate the impact of variables like price changes, economic indicators, or marketing efforts on sales.

3. Implemented approach or Algorithm:

Algorithms Explanation:

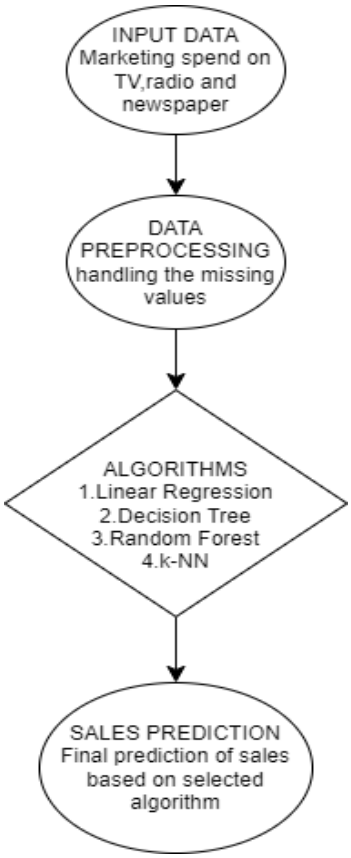
- **Linear Regression** finds the best-fitting straight line that predicts sales based on a single factor (like TV spend). It estimates how much sales change with each unit increase in the factor.
- **Decision Trees** split the data into branches based on different features, creating rules to predict sales. Each split is made to reduce error, and the final prediction is the average value of the target variable in the leaf node.
- **Random Forest** builds many decision trees on random subsets of the data and averages their predictions. This reduces overfitting and improves accuracy.
- **k-Nearest Neighbors (k-NN)** makes predictions by looking at the K closest data points in the dataset. It averages their sales values to predict the sales for a new point, based on similarity.

Justification for Selection: Using a variety of algorithms for sales prediction allows for a comprehensive analysis of different data characteristics and relationships. **Linear Regression** is ideal for simple, linear relationships with a single predictor. **Decision Trees** excel in handling complex, non-linear relationships and offer easy interpretability, whereas **Random Forest** enhances accuracy and reduces overfitting by averaging predictions from multiple trees. Lastly, **k-Nearest Neighbors (k-NN)** captures local patterns by predicting sales based on the closest previous instances. This combination of algorithms ensures a reliable approach to accurately predicting sales across various scenarios.

Working: The implemented algorithm follows these steps:

1. **Data Collection:** Gather data from the Kaggle dataset on advertising expenditures and sales.
2. **Data Preprocessing:**
    - Check for Missing Values: Identify and handle any missing data.
    - Check for Duplicate Values: Remove duplicates to ensure data integrity.
    - Check for Outliers: Identify extreme values that could skew results.
3. **Exploratory Data Analysis (EDA):**
    - Visualize the distribution of sales.
    - Analyze relationships between sales and independent variables.
4. **Model Building:** Train the models on the data.
5. **Prediction:** Use the trained models to predict sales based on marketing spend.

FLOWCHART:

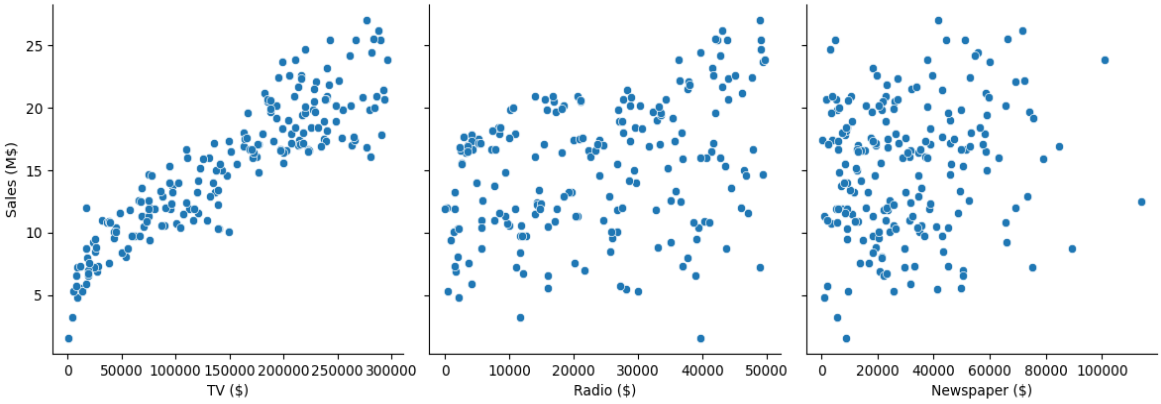


Working Flowchart

4. Results:

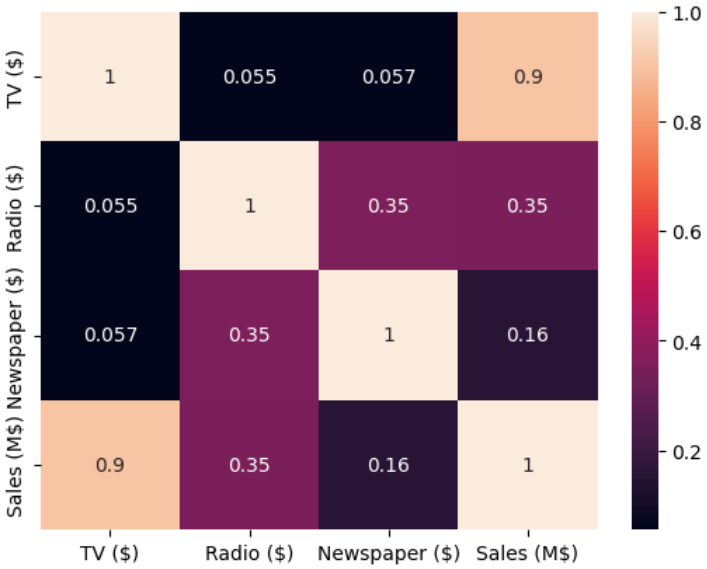
The models were evaluated based on metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared value.

Figure 1: Distribution of Sales



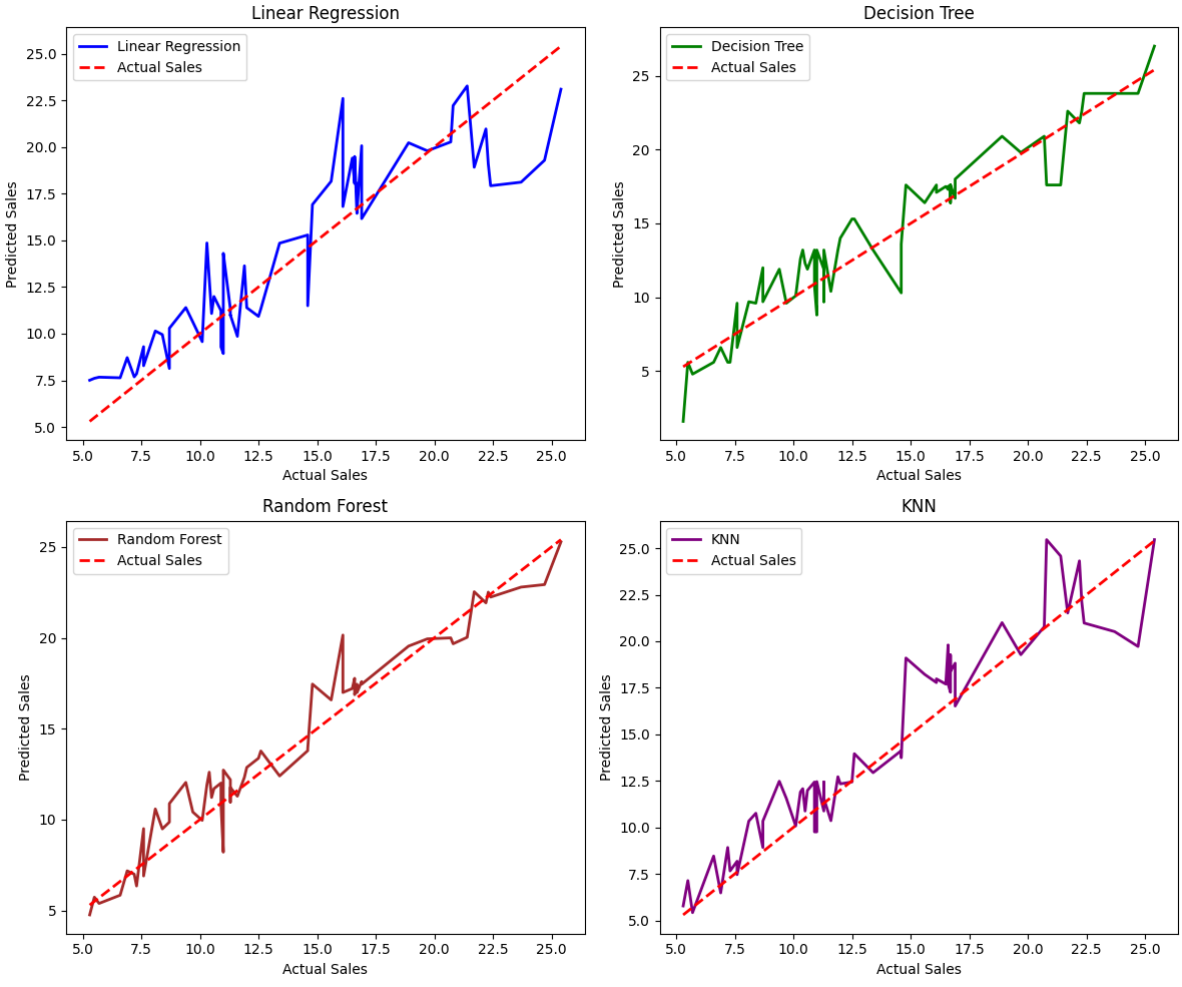
The plot illustrates the distribution of the sales variable with respect to investments on different platforms.

Figure 2: Correlation Heatmap



The heatmap indicates strong positive correlations between sales and marketing expenditures, particularly for TV advertising.

Figure 3: Model Performance



This plot compares the actual vs. predicted sales for various models, showing the accuracy of predictions.

Models	R-squared Score
1. Linear Regression	81.48
2. Decision Tree	89.02
3. K-Nearest Neighbors	88.31
4. Random Forest	94.34

Results Explanation: The results indicate that models successfully predict sales, where Random Forest algorithm was able to achieve the highest accuracy of 94.34. The correlation analysis confirmed that increased advertising spend on TV positively influences sales more than other platforms.

5. Conclusion:

In conclusion, this sales prediction project effectively applied various algorithms to predict sales based on marketing spend across platforms like TV, radio, and newspapers. **Linear Regression** handled simple, linear relationships, while **Decision Trees**, **Random Forest**, and **k-NN** captured more complex, non-linear patterns. Comparing these algorithms allowed us to choose the most suitable model for accurate predictions. The project demonstrates the importance of using diverse approaches to improve sales **prediction** and optimize marketing strategies.

6. References:

- Kaggle Advertising Dataset. Retrieved from Kaggle.
- Documentation of Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn libraries.