

# Sales Prediction Using Python

## Model Development & Deployment Documentation

---

### 1. Project Overview

This project focuses on predicting product sales based on advertising expenditure across multiple marketing channels using machine learning techniques in Python. The objective is to analyze historical advertising data, build a regression model, and deploy the trained model as an interactive Streamlit web application.

The project demonstrates an end-to-end data science workflow, covering data analysis, feature engineering, model training, evaluation, and deployment.

The work is divided into two major components:

1. Data Analysis & Model Development (Jupyter Notebook)
  2. Model Deployment using Streamlit
- 

### 2. Dataset Description

The dataset used in this project is the Advertising Dataset, which is commonly used for regression analysis and sales prediction tasks.

#### Dataset Source:

Kaggle – Advertising Dataset

#### Dataset Features:

- **TV** – Advertising budget spent on TV promotions
- **Radio** – Advertising budget spent on radio promotions
- **Newspaper** – Advertising budget spent on newspaper promotions
- **Sales** – Product sales (target variable)

The dataset is fully numerical and contains no missing values, making it suitable for regression-based machine learning models.

---

### 3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand feature distributions and their relationships with the target variable (Sales).

## **Steps Performed:**

- Loaded the dataset and examined shape, data types, and summary statistics
- Analyzed the distribution of the target variable (Sales)
- Performed univariate analysis on advertising spend features
- Identified outliers using boxplots
- Conducted bivariate analysis using scatter plots
- Performed correlation analysis using a heatmap
- Used pair plots to study multivariate relationships

## **Key EDA Findings:**

- TV advertising has a strong positive relationship with Sales
  - Radio advertising has a moderate influence on Sales
  - Newspaper advertising shows weak correlation with Sales
  - Sales distribution is approximately linear
  - Linear Regression is appropriate for this prediction task
- 

## **4. Data Cleaning & Feature Engineering**

To prepare the data for modeling, the following preprocessing steps were performed:

- Verified that the dataset contains no missing values
- Ensured all features were numerical and suitable for regression
- Selected relevant features based on correlation and EDA insights
- Ensured consistency between training and deployment input features

The final dataset was kept simple to maintain model interpretability and deployment compatibility.

---

## **5. Model Training**

The problem was formulated as a supervised regression task.

### **Model Used:**

- **Linear Regression**

### **Training Process:**

- Split the dataset into training and testing sets
- Trained the Linear Regression model using Scikit-learn
- Evaluated performance on unseen test data
- Saved the trained model using `joblib` for deployment

The trained model was stored as `lr_model.pkl`.

---

## 6. Model Evaluation

Model performance was evaluated using regression metrics.

### Evaluation Metrics:

- $R^2$  Score
- Mean Squared Error (MSE)

### Evaluation Results:

- High  $R^2$  score indicating good predictive accuracy
  - Low error values showing reliable sales predictions
  - The model generalized well to test data
- 

## 7. Streamlit Application Integration

The trained Linear Regression model was integrated into an interactive Streamlit web application.

### Application Workflow:

1. Load the trained model (`lr_model.pkl`)
2. Accept user inputs:
  - o TV Advertising Spend
  - o Radio Advertising Spend
  - o Newspaper Advertising Spend
3. Convert user input into a NumPy array
4. Predict sales using the trained model
5. Display predicted sales value on the UI

The application provides a simple and intuitive interface for real-time predictions.

---

## 8. Deployment Using Streamlit

The Streamlit application was deployed to enable web-based access without requiring local setup.

The deployment demonstrates hands-on experience in:

- Model serialization using joblib
  - Building interactive ML applications
  - Deploying machine learning models using Streamlit
- 

## 9. Project Structure

```
Sales-Prediction-Using-Python/
├── SALES_PREDICTION_USING_PYTHON.ipynb
│   └── Data analysis, EDA, model training & evaluation
├── Advertising Dataset
│   └── Dataset used for training and testing
├── lr_model.pkl
│   └── Trained Linear Regression model
├── app.py
│   └── Streamlit application
├── requirements.txt
│   └── Python dependencies
├── sales_predictions.csv
│   └── Sample prediction outputs
└── README.md
    └── Project documentation
```

---

## 10. Key Learnings

- Understanding regression-based machine learning problems
  - Performing structured Exploratory Data Analysis (EDA)
  - Training and evaluating regression models
  - Model serialization using joblib
  - Deploying machine learning models using Streamlit
  - Building end-to-end data science projects
- 

## 11. Conclusion

This project demonstrates a complete machine learning pipeline for sales prediction, starting from data analysis and model development to deploying a production-ready web application. The integration of EDA, regression modeling, and Streamlit deployment reflects practical data science and machine learning skills applicable to real-world business problems.

---

# **Author**

**Durga Prasad**

GitHub: <https://github.com/shettyprasad-git>

LinkedIn: <https://www.linkedin.com/in/durgaprasadshetty>