SALES PREDICTION USING PYTHON

PROJECT INTRODUCTION

This project uses Python-based machine learning techniques to predict **product sales** based on advertising spend across **TV**, **Radio**, and **Newspaper** platforms. It demonstrates a complete workflow from data preprocessing to model evaluation, with a focus on business impact.

🎯BUSINESS PROBLEM AND OBJECTIVE

BUSINESS PROBLEM:

Businesses spend large amounts on ads, but not all channels yield the same ROI. Understanding which channel impacts sales the most is critical.

OBJECTIVE:

- Predict product sales based on advertising spend
- Identify which advertising mediums drive higher sales
- Build a reliable, interpretable, and optimized regression model

DATASET INFORMATION

SOURCE: Sales_Data.csv (200 rows × 4 columns)

📌**FEATURE DESCRIPTIONS**:

FEATURE	TYPE	DESCRIPTION
TV	Float64	Advertising spend on TV (in thousands)
Radio	Float64	Advertising spend on Radio (in thousands)
Newspaper	Float64	Advertising spend on Newspaper (in thousands)
Sales	Float64	Units sold (Target variable)

🎯 **TARGET VARIABLE:**

- Sales (Continuous numerical value)

🧹DATA PREPROCESSING

HANDLING MISSING DATA:

- No missing values found in the dataset.

ENCODING / SCALING:

- Applied StandardScaler for feature normalization
- Used LabelEncoder where applicable (though minimal due to numeric features)

DATA TYPES & CONVERSIONS:

- Verified all columns are in appropriate numerical format (float64)

📊EXPLORATORY DATA ANALYSIS (EDA)

📈 VISUALIZATIONS USED:

- Pair Plot
- Correlation Heatmap
- Regression Plot

- Feature Distribution Plots

TRENDS & OBSERVATIONS:

- Strong positive correlation between **TV** and **Sales**
- **Radio** also positively impacts Sales
- **Newspaper** has a weak influence
- No major multicollinearity detected

🧠FEATURE ENGINEERING

NEW FEATURES CREATED:

- No new features were engineered; original features were sufficient.

FEATURE SELECTION INSIGHTS:

- All three features were retained initially
- Later insights suggest Newspaper may be excluded due to low impact

MODELING

ALGORITHMS USED:

- Linear Regression (Baseline)
- Random Forest Regressor
- XGBoost Regressor (Final model)

FINAL MODEL CHOICE:

- **XGBoost Regressor** provided the best performance

HYPERPARAMETER TUNING:

- Performed basic tuning using default XGBoost parameters
- Applied cross_val_score for validation

EVALUATION

METRICS USED:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- R^2 Score

PERFORMANCE SUMMARY:

MODEL	R^2 SCORE	MAE	MSE
Linear Regression	~0.90	Moderate	Moderate
Random Forest	~0.94	Lower	Lower
✅ **XGBoost (Final)**	**~0.96**	Lowest	Lowest

VISUAL COMPARISONS:

- Predicted vs. Actual plots showed high alignment, especially with XGBoost

KEY RESULTS AND BUSINESS INSIGHTS

- **TV and Radio** are the most effective ad channels
- **Newspaper** contribution to sales is minimal
- Predictive modeling can optimize ad spend allocation
- The model is robust with strong predictive accuracy (R^2 ~ 0.96)

CHALLENGES ENCOUNTERED AND SOLUTIONS

CHALLENGE	SOLUTION
Small dataset	Used cross-validation to ensure model generalization
Weak importance of Newspaper	Kept initially; evaluated during model comparison
Scaling needs for some models	Applied StandardScaler to normalize features

TOOLS, LIBRARIES, AND FRAMEWORKS USED

- 🐍**Python 3.x**

- 📊**Pandas, NumPy** – Data loading and preprocessing
- 📉**Matplotlib, Seaborn** – Visualizations
- 🧠**Scikit-learn** – Modeling, splitting, evaluation
- ⚡**XGBoost** – Final model

FINAL THOUGHTS

This project is a practical demonstration of how regression models can empower businesses with **data-driven decisions** in ad budget planning. It also reinforces the importance of proper EDA, preprocessing, and model evaluation.

👤ABOUT ME

Saran S

EMAIL: saranselvaraj2401s@gmail.com
LINKEDIN: linkedin.com/in/saranselvaraj2401
GITHUB: github.com/saran2007s