

ADVERTISING IMPACT ON SALES PREDICTION 🚀

In this project, I developed a predictive model to estimate sales based on advertising spending across various media channels (TV, Radio, and Newspaper). This project employed robust data science methodologies, leveraging multiple libraries and techniques to derive meaningful insights and predictions.

1. Data Importation 📥

I initiated the project by importing essential libraries, including:

- Pandas for data manipulation and analysis.
- NumPy for numerical computations.
- Matplotlib and Seaborn for data visualisation.
- Scikit-learn for machine learning model implementation.

2. Dataset Loading 📊

I loaded the dataset using Pandas and displayed the first few rows to understand its structure. This dataset contains 200 entries and four columns: TV, Radio, Newspaper, and Sales.

3. Data Cleaning and Exploration 🔍

I conducted a thorough examination of the dataset to identify any inconsistencies:

- Reviewed the dataset info and descriptive statistics to understand the data distribution.
- Assessed for missing values and handled them by dropping rows or filling them with mean values.
- I created a correlation matrix visualisation to identify relationships between variables, revealing that advertising spend significantly correlates with sales.

4. Feature Selection and Scaling 🔧

I selected the features (TV, Radio, and Newspaper) that would contribute to the sales prediction. To ensure that the model would not be biased towards any particular scale, I applied StandardScaler to standardise the feature values.

5. Data Splitting 📁

I split the dataset into training (80%) and testing (20%) sets to validate the model's performance accurately.

6. Model Initialization and Training 🏋️

I employed the Random Forest Regressor, a robust ensemble learning method, to predict sales. The model was trained using the training dataset, allowing it to learn from the feature relationships.

7. Predictions and Evaluations 📈

Upon training, I made predictions on the test set and evaluated the model's performance using key metrics:

- Root Mean Squared Error (RMSE): 0.77, indicating the average prediction error.
- R-squared (R^2): 0.98, suggesting that the model explains 98% of the variance in sales.

These results demonstrate the model's high accuracy and reliability in predicting sales based on advertising expenditure.

8. Results Visualization 📊

To enhance the interpretability of the model results, I generated several plots:

- Actual vs. Predicted Sales: Most data points fell close to the ideal prediction line, indicating strong predictive performance.
- Residual Plot: The residuals were randomly distributed around the horizontal line, confirming that the model's predictions were unbiased.
- Feature Importance Plot: This revealed that TV advertising had the highest impact on sales, followed by Radio and Newspaper.
- Learning Curves: The training and cross-validation scores were close, indicating that the model is not overfitting and generalises well to unseen data.

9. Skills Developed 🧠

Throughout this project, I enhanced my skills in data cleaning, exploratory data analysis (EDA), model evaluation, feature engineering, and data visualisation. The hands-on experience with libraries such as Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn was invaluable.

Key Takeaways 🎯

- The project successfully predicted sales with high accuracy using Random Forest regression.
- Insights gathered can inform strategic advertising decisions, optimising marketing budgets across channels.

Hashtags:

#DataScience #MachineLearning #Python #SalesPrediction #RandomForest
#DataVisualization #Pandas #NumPy #ScikitLearn #FeatureEngineering
#ModelEvaluation #PredictiveAnalytics #DataCleaning #ExploratoryDataAnalysis
#LearningCurves #ArtificialIntelligence #DataDriven #DataAnalysis #Statistics
#Analytics