Data Analysis Report: Sales Prediction

Objective

The primary objective of this analysis was to predict sales based on advertising spend across different channels (TV, Radio, Newspaper) and evaluate the performance of models like **Linear Regression** and **Random Forest**. Key metrics such as **Mean Squared Error (MSE)** and **R-squared** were used to assess model accuracy.

Models and Metrics

A. Linear Regression

- **Mean Squared Error (MSE):** 2.9078
    - This indicates that the average squared difference between actual and predicted sales is about 2.91 units.
- **R-squared:** 0.9059
    - This suggests that **90.59%** of the variance in sales is explained by the model.

B. Random Forest

- **Mean Squared Error (MSE):** 1.4374
    - The Random Forest model has significantly lower error than Linear Regression, showing better predictive performance.
- **R-squared:** 0.9535
    - This indicates that **95.35%** of the variance in sales is explained by the Random Forest model.

C. Cross-Validated R-squared Scores

- The cross-validation process yielded R-squared scores ranging from **0.8455 to 0.9318**, with an average R-squared of **0.8954**.
    - This shows the Random Forest model's performance is consistent across multiple train-test splits.

Feature Analysis

From the Linear Regression coefficients:

- **Radio advertising** had the most significant impact on sales, with a coefficient of **0.1009**.
- **TV advertising** had a moderate impact, with a coefficient of **0.0545**.
- **Newspaper advertising** had the least impact, with a coefficient of **0.0043**.

The Random Forest model likely captured non-linear relationships, improving accuracy compared to the simpler Linear Regression model.

Insights

1. **Model Performance**:
    - Random Forest outperformed Linear Regression in terms of both MSE and R-squared. It captured complex interactions in the data that Linear Regression couldn't model effectively.
    - However, Linear Regression still provides interpretable results and is useful for understanding feature importance.
2. **Advertising Strategy**:
    - **Radio advertising** offers the best ROI, as it has the most significant positive impact on sales.
    - While **TV advertising** also contributes, its impact is smaller compared to Radio.
    - **Newspaper advertising** shows minimal contribution to sales and may not justify significant investment.
3. **Cross-Validation**:
    - The consistent R-squared scores during cross-validation highlight the Random Forest model's robustness and reliability for making predictions.

Recommendations

1. **Model Usage**:
    - Use the **Random Forest model** for operational sales predictions, as it is more accurate.
    - Retain Linear Regression for quick insights into feature importance and interpretability.
2. **Advertising Budget Allocation**:
    - Increase spending on **Radio** campaigns to maximize sales impact.
    - Maintain a moderate investment in **TV** advertising.
    - Consider reallocating funds from **Newspaper advertising** to other channels for better ROI.
3. **Future Analysis**:
    - Experiment with other advanced models (e.g., Gradient Boosting, XGBoost) to potentially improve prediction accuracy.

- Conduct **time-series analysis** if sales data includes temporal patterns, such as seasonal fluctuations.

4. **Data Enhancement**:
    - Gather more granular data, such as **customer demographics** or **regional sales trends**, to refine predictions.
    - Include other features like **competitor activity**, **online vs. offline spend**, or **holiday seasons** for a more comprehensive analysis.

Conclusion

The Random Forest model is the preferred choice due to its superior accuracy and robustness. By optimizing advertising strategies based on these insights, businesses can effectively allocate resources and maximize sales. Future efforts should focus on expanding the dataset and incorporating additional predictive features.

****