Data Analysis Report: Sales Data Analysis - Linh Nguyen

I. Introduction

Understanding consumer behavior is the key to success in the world of business. In this report, we analyze sales data of a product, focusing on understanding the impact of various marketing strategies and promotional activities such as radio advertisements, in-store promotions, discounts, television campaigns, stock availability, and online advertising on sales volume. By employing advanced analytical techniques, including Multiple Linear Regression and Decision Tree Regression, we aim to decode the underlying dynamics of consumer purchasing behavior and empower businesses with actionable insights.

The report is divided into 4 distinct sections. The first section, **Data**, provides a detailed overview of the dataset and the Exploratory Data Analysis (EDA), providing insights into the relationships among variables employing statistical measures and visualizations methods. The next section is **Analysis/ Method** which explains the results and performance of advanced techniques like Multiple Linear Regression and Decision Tree Regression on predicting sales. The **Conclusion** section will summarize key findings, highlighting the most effective marketing strategies and providing actionable recommendations for businesses. The last section, **Appendix**, will provide the code that I used for the analysis.

II. Data

1. Data Overview

The sales dataset is from Kaggle, which includes information on the sales of a product in two years. Each row in the dataset represents the sales volume for a week, along with details about the marketing campaigns and promotional methods used for the product.

2. Data Dictionary

- Sale: This variable contains numerical data representing the number of product sales for each observed week.
- Price: The observed week's base price for the product.
- Radio: The number of radio advertisements or campaigns promoting the product for the observed week.
- InStrSpending: The average expenses associated with promoting the product in stores for the observed week.
- Discount: The discount rate applicable for the observed week.
- TVSpending: The average expenditure on television campaigns during the observed week.
- StockRate: The stock-out rate, calculated as the number of times the product was out of stock divided by the total number of product visits.

- OnlineAdsSpending: The online ads spending, calculated the total amount of spend on online advertising.

3. Exploratory Data Analysis

3.1. Univariate Analysis

Table 1 provides descriptive statistics for all variables in the dataset, offering insights into their distributions. Figure 1 visually represents these distributions, highlighting that the 'Sale' variable adheres to a normal distribution, with a mean and median both around $170,000, demonstrated by its bell-shaped curve. In contrast, other variables exhibit more uniform distributions. Notably, the absence of outliers in these variables is evident from the histograms.

	count	mean	std	min	25%	50%	75%	max
Sale	992.000	171,327.119	81,397.843	1,992.000	112,479.250	170,390.500	226,027.250	393,914.000
InStrSpending	992.000	30.593	17.493	0.190	14.830	31.385	45.660	59.960
Discount	992.000	0.251	0.145	0.000	0.130	0.250	0.380	0.500
TVSpending	992.000	98.679	57.117	0.130	49.638	97.510	147.620	199.910
StockRate	992.000	0.495	0.287	0.000	0.250	0.490	0.740	1.000
Price	992.000	14.600	8.716	0.140	6.917	14.820	22.100	29.990
Radio	992.000	1,479.570	885.420	4.000	708.250	1,413.500	2,273.000	2,997.000
OnlineAdsSpending	992.000	1,596.504	927.475	12.540	786.327	1,595.455	2,420.688	3,198.270

Table 1: Descriptive statistics of all variables in the dataset

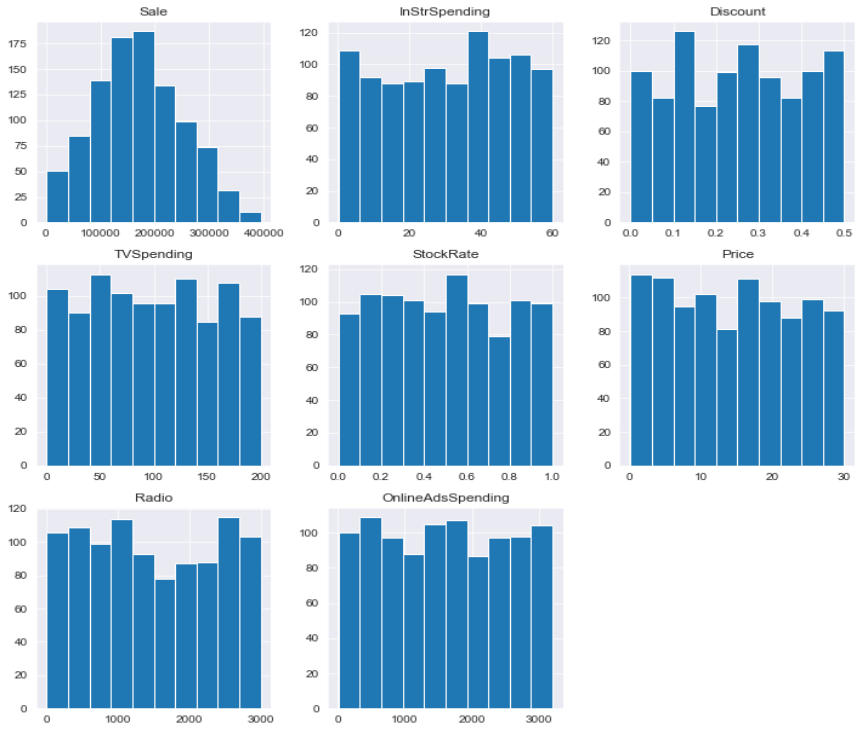

Figure 1: Histogram of all variables in the dataset

3.2. Bivariate Analysis

a. Effect of Marketing Strategies on each other

From Figure 2, it becomes apparent that each marketing strategy shows no significant impact on the others, as indicated by a correlation coefficient of 0. This is critical in building the Multiple Linear Regression model later in section 2. In cases where variables are highly correlated, it can lead to multicollinearity issues, making it challenging to discern the individual effect of each variable on the outcome. By ensuring low intercorrelation among marketing strategies, the model can accurately identify and quantify the unique contribution of each strategy to the sales outcome, facilitating more precise predictions and informed business decisions.

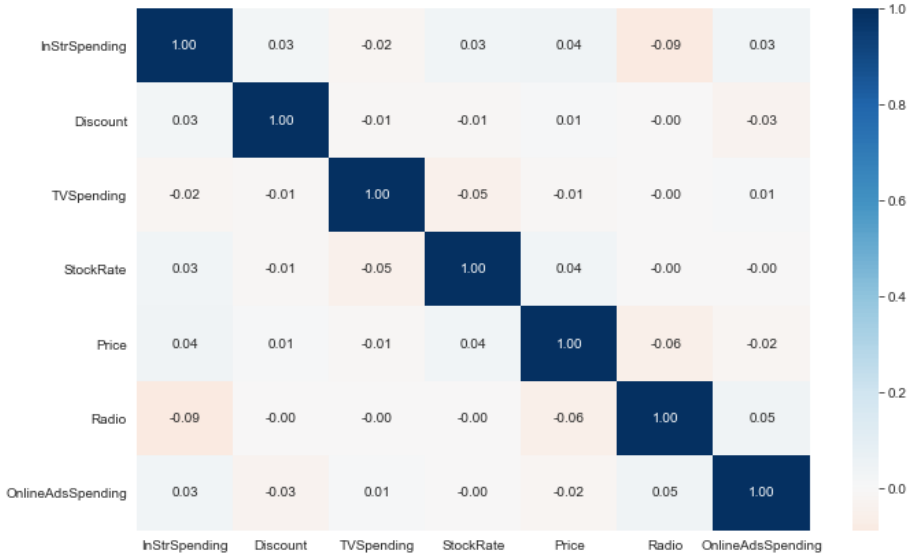

Figure 2: Correlation matrix among the variables (not including Sale variable)

b. Effect of each Marketing Strategies on Sale

Figure 3 provides insights into the relationship between specific marketing strategies and sales. From Figure 3, we can see that Price, TVSpending, and In-store Spending have a relatively strong linear relationship with Sale with a correlation of -0.67, 0.41, and 0.51 respectively, and Figure 4 is the scatterplots among these variables. This indicates that an increase in the product's price leads to a decrease in sales, while higher investments in TV and in-store promotions positively influence sales. Based on these findings, it is advisable for the business to focus on enhancing in-store and TV promotional campaigns, while also considering a strategic reduction in the product's price to boost overall sales.

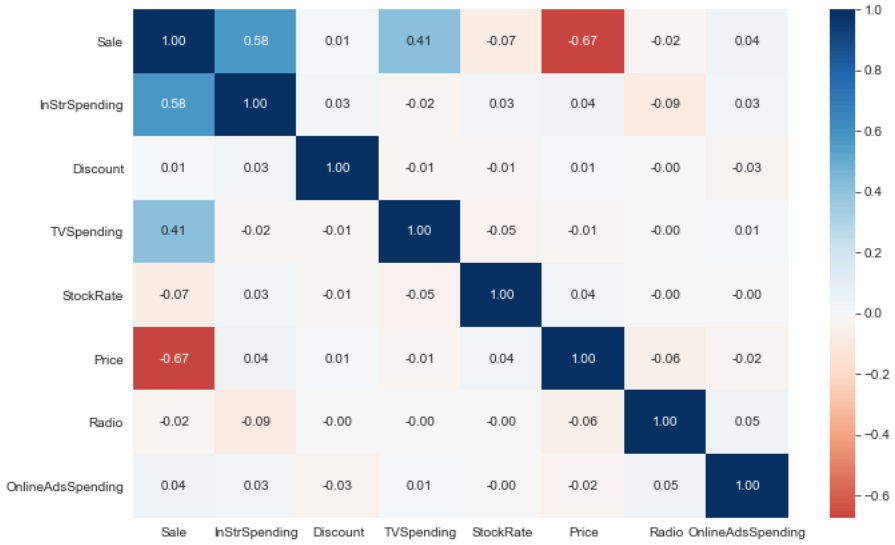

Figure 3: Correlation matrix among the variables (including Sale variable)

Figure 4: Scatterplot showing relationship between In-store Spending, Price, and TV Spending vs Sale

III. Analysis/ Method

In this data analysis report, our focus is to explore how each marketing strategy influences product sales. We examined several key variables, including 'InStrSpending', 'Discount', 'TVSpending', 'StockRate', 'Price', 'Radio', and 'OnlineAdsSpending', to understand their individual and collective impact on sales volume ('Sale').

Before developing predictive models, we first divide our dataset into two distinct sets: a training dataset comprising 60% of the data and a validation dataset consisting of the remaining 40%.

This allows us to develop and fine-tune our predictive model on a substantial portion of the data, ensuring its accuracy and reliability.

1. Linear Regression

1.1. Multiple Linear Regression Model

Figure 2 shows the performance of Multiple Linear Regression on both training and validation data on different metrics. With the coefficient of determination of 0.99, it shows that the model performs very well on explaining most of the variance in sales. With low Mean Absolute Percentage Errors (MAPE) of 3.75% for training data and 4.98% for validation data, the model's predictions are close to actual values. Additionally, the Root Mean Squared Error (RMSE) is small, indicating accurate predictions. Although there is a slight increase in error metrics for the validation data, the model still performs well in capturing the relationships between variables and sales volume.

Data	Mean Absolute Percentage Error (MAPE)	Root Mean Squared Error (RMSE)	Coefficient of Determination (R^2)
Training	3.75%	3125.44	0.999
Validation	4.98%	3155.27	0.998

Table 2: Performance Summary of Multiple Linear Regression Model

Table 3 shows the importance of each marketing strategy on sales volume. Based on these coefficients, we can see that stock-out rate (StockRate) and Price have the most significant impact on sales. The negative relationships suggest that when the product frequently runs out of stock, it adversely affects customer satisfaction and, consequently, sales and customers tend to purchase more when the product is priced lower, highlighting the price sensitivity of consumers. These findings emphasize the critical importance of managing stock availability and pricing strategies. Ensuring products are consistently in stock and competitively priced can significantly enhance sales performance.

Predictor	Coefficient
InStrSpending	2882.179
Discount	4425.186
TVSpending	590.297

StockRate	-13609.314
Price	-6487.981
Radio	0.00368
OnlineAdsSpending	0.123

Table 3: The importance of each marketing strategy on sales volume for Multiple Linear Regression Model

1.2. Linear Regression with Lasso CV

Based on our Bivariate Analysis, we can see that only Price, TVSpending, and In-store Spending have significant influence on our target variable, Sale. Thus, we use Lasso Cross-Validation (LassoCV) to select important variables and prevent overfitting by developing a simpler and more interpretable model.

As shown in Table 4, comparing the performance metrics of the model after LassoCV regularization to the original model, it is evident that LassoCV has improved the model's performance. In the training data, both models exhibit similar results, with nearly identical Mean Absolute Percentage Error (MAPE). The improvement is most apparent in the validation data. The LassoCV model demonstrates a reduction in Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) with a value of 3139.80 and 4.87% respectively compared to a RME of 3155.27 and a MAPE of 4.98% in the original model, suggesting a better accuracy in predicting sales. The R-squared value for the validation data remains high for both models, indicating that the LassoCV model can still explain a significant portion of the variance in the data while improving the accuracy in its prediction.

In short, the LassoCV model outperforms the original model by enhancing accuracy and improving precision, making it a more reliable predictor for future sales.

Data	**Mean Absolute Percentage Error (MAPE)**	**Root Mean Squared Error (RMSE)**	**Coefficient of Determination (R^2)**
Training	3.70%	3128.42	0.999
Validation	4.87%	3139.80	0.998

Table 4: Performance Summary of Multiple Linear Regression Model after Lasso Cross-Validation

After applying LassoCV, the coefficients for the predictors have undergone significant changes, indicating the regularization effect of the Lasso technique on the model. From Table 5, the most significant marketing strategies that impact sales volume are Price, TV marketing campaigns (TVSpending) and in-store marketing campaigns (InStrSpending), which are consistent with our findings in the univariate analysis. The coefficient for Radio is shrunk to zero, suggesting that radio advertising does not significantly impact sales and becomes negligible in influencing sales volume. Comparing these coefficients with the original model, it is evident that LassoCV has regularized the model, shrinking the coefficients toward zero. This regularization helps prevent overfitting and promotes a more generalized model. While some variables have experienced a reduction in their impact, the overall trends in how these marketing strategies influence sales have been preserved, allowing for a more stable and reliable predictive model.

Predictor	**Coefficient**
InStrSpending	50,068.674
Discount	590.745
TVSpending	33,422.491
StockRate	-3,921.518
Price	-57,054.024
Radio	0.000
OnlineAdsSpending	56.828

Table 5: The importance of each marketing strategy on sales volume for Multiple Linear Regression Model after Lasso Cross-Validation

2. Decision Tree Regressor

Multiple Linear Regression is used when there is a linear relationship among predictors and dependent variable. However, what if these marketing strategies do not influence sales volume linearly? In this case, we develop Decision Tree Regressor with the aim to comprehend the relationship between marketing strategies and sales when the interactions among these marketing strategies influence sales differently.

2.1. Decision Tree Regressor

After training the Decision Tree Regressor with max_depth = 3, the model's performance is summarized in Table 6. Comparing the training and validation performance, the increase in the error metrics in the validation data suggests that the model might generalize slightly less

effectively to unseen data. Furthermore, the Decision Tree Model does not perform as well as compared with the Multiple Linear Regression Model with R-squared around 0.74, indicating that the model can only explain 74% of the variance in the sales data. This sign of overfitting suggests that further model tuning might be necessary to address these issues and enhance the model's performance.

Data	Mean Absolute Percentage Error (MAPE)	Root Mean Squared Error (RMSE)	Coefficient of Determination (R^2)
Training	31.74%	38461.55	0.77
Validation	39.64%	40962.18	0.74

Table 6: Performance Summary of Decision Tree Regressor Model

2.2. Decision Tree Regressor after Hyperparameter Tuning

After tuning the Decision Tree Regressor Model (see Appendix), there is a notable improvement in performance metrics such as Mean Absolute Error (MAPE) and Root Mean Squared Error (RMSE) for the validation data as shown in Table 7. The MAPE and RMSE are halved, indicating a significant reduction in prediction errors. Additionally, the R-squared value increases substantially from 0.74 to 0.92, suggesting that the model explains 92% of the variance in the validation data.

However, this enhanced regression model displays signs of high variance. It achieves an exceptionally high R-squared score (0.99) on the training data, along with very low MAPE (4.42%) and RSME (6890.75), indicating an outstanding fit. These low errors and high R-squared suggest the model captures the noise of the training data very well. Meanwhile, this precision comes at a cost: the model's performance on the validation data is significantly reduced. The substantial difference between training and validation metrics indicates a potential overfitting issue.

Data	Mean Absolute Percentage Error (MAPE)	Root Mean Squared Error (RMSE)	Coefficient of Determination (R^2)
Training	4.42%	6890.75	0.99
Validation	17.73%	23282.09	0.92

Table 7: Performance Summary of Hyperparameter-tuned Decision Tree Regressor Model

3. Model Comparison

Based on the metrics presented in Table 8, it is evident that the performance of the best Decision Tree Regressor model falls short in comparison to the best Multiple Linear Regression model. The Decision Tree model shows higher errors and lower coefficients of determination (R-squared values) in the validation dataset when compared to the Multiple Linear Regression model. These metrics indicate that the relationship between marketing strategies lean more towards linear so that the Multiple Linear Regression model is more appropriate to predict the sale volume. Furthermore, the Multiple Linear Regression Model is more interpretable with the coefficients. This interpretability of the Multiple Linear Regression model is a crucial advantage for businesses. The coefficients associated with each marketing strategy offer clear insights into their impact on sales volume. This transparency allows marketing teams to make well-informed decisions, adjusting strategies based on the quantified influence of each variable.

In summary, while Decision Tree Regressor is a powerful tool for exploring complex relationships, the dataset's underlying linear nature and the need for interpretability make Multiple Linear Regression, especially after Lasso Cross-Validation, the more suitable choice for predicting sales volume in this context. It not only provides accurate predictions but also equips businesses with actionable insights, enabling them to refine marketing strategies effectively and drive sustainable growth.

	Hyperparameter-Tuned Decision Tree Regressor (Validation)	Multiple Linear Regression LassoCV (Validation)
Mean Absolute Percentage Error (MAPE)	17.73%	4.87%
Root Mean Squared Error (RMSE)	23282.09	3139.80

Coefficient of Determination (R^2)	0.92	0.998

Table 8: Comparison between the performance of Decision Tree Regressor after fine-tuned and Multiple Linear Regression after Lasso Cross Validation

IV. Conclusion

In this sales data analysis, we explored the intricate relationship between various marketing strategies and sales volume by examining the impact of spending on radio advertisements, in-store promotions, discounts, television campaigns, stock availability, and online advertising on sale volume. Through the Exploratory Data Analysis, we can see that TVSpending, Price, and InStrSpending have a strong relationship with sales volume, indicating their significant influence on customer purchasing behavior. The correlation analysis revealed that Price and InStrSpending had negative correlations with Sale, implying that higher prices and in-store spending led to decreased sales. In contrast, TVSpending had a positive correlation, suggesting that increased investment in television campaigns positively affected sales.

To predict sales volume, we employed two major techniques: Multiple Linear Regression and Decision Tree Regressor. Multiple Linear Regression allowed us to quantify the impact of each marketing strategy accurately. The model, after Lasso Cross-Validation, highlighted the crucial roles of Price, TVSpending, and InStrSpending, shedding light on the necessity of strategic pricing and effective television and in-store promotions.

Additionally, we explored the complex interactions between marketing strategies using Decision Tree Regressor. Despite its ability to capture intricate patterns, the Decision Tree model struggled to generalize well to unseen data, evident from its higher errors and lower R-squared values in the validation dataset. This suggested that the relationship between marketing strategies leaned more towards linearity, favoring the Multiple Linear Regression approach.

In conclusion, this analysis not only provided valuable insights into the impact of different marketing strategies on sales volume but also emphasized the importance of selecting appropriate modeling techniques. The Multiple Linear Regression model, with its interpretability and accuracy, proved to be the optimal choice for understanding and predicting sales based on the given marketing variables. Businesses can leverage these findings to optimize their marketing campaigns, pricing strategies, and promotional activities, ultimately enhancing their sales performance and customer satisfaction.