**Title :** An Analysis on Optimizing Coupon Distribution Based on Spending Behaviour
Author : Pratik Ganguli (pgan501)

The **Salmon Stores** dataset captures customer spending behavior and promotional response data for 1,000 customers of a national women's apparel retail chain. The company aims to optimize its marketing strategy by identifying customers who are most likely to use discount coupons. This analysis examines the relationship between **Spending**, a continuous independent variable representing last year's total spending, and **Coupon** usage, a binary dependent variable indicating whether a customer redeemed a previous promotional coupon. To achieve this, a logistic regression model is applied to predict which customers are most likely to redeem a coupon. These insights will help Salmon Stores send catalogs only to customers with the highest probability of using the coupon, ensuring a more targeted and cost-effective promotional strategy.
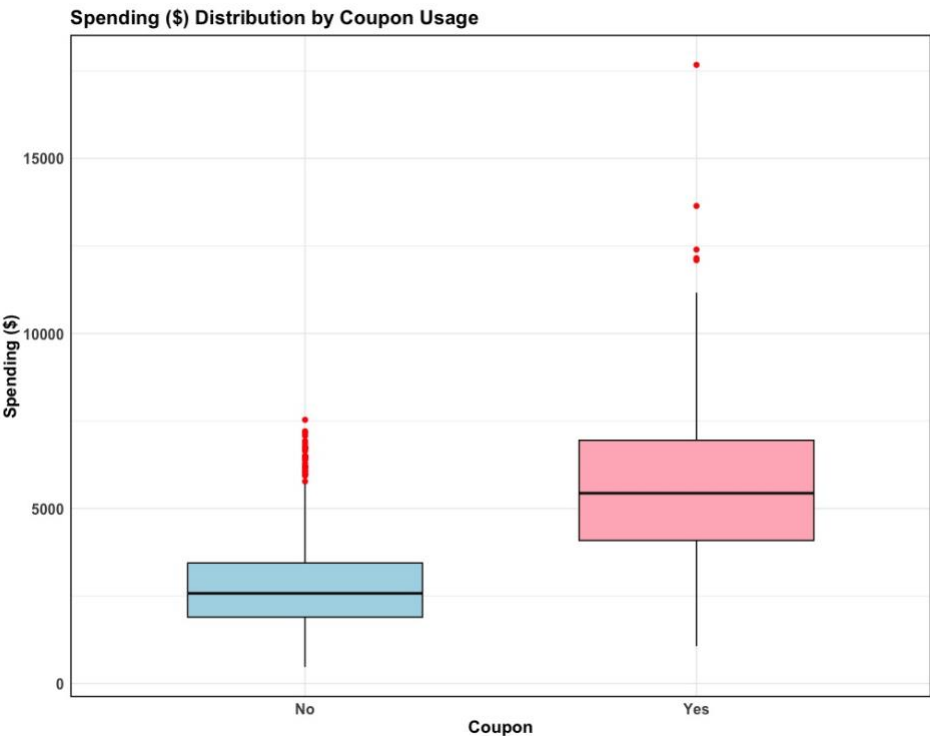

Figure 1.1 *Spending Distribution by Coupon Usage (Boxplot)*

Spending	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Coupon(YES)	1068	4087	5438	5744	6949	17675
Coupon(NO)	472	1894	2576	2821	3445	7534

Table 1.1 *Summary Statistics of Spending by Coupon Usage (unit = $)*

The summary statistics[Table 1.1] indicate that customers who used a coupon tend to have significantly higher spending than those who did not. The median spending for coupon users is $5,438, compared to $2,576 for non-users, suggesting a strong association between higher spending and coupon redemption. Additionally, the interquartile range(IQR) for coupon users($4,087–$6,949) is notably wider than that of non-users($1,894–$3,445), highlighting greater variability in spending behaviour among those who redeem coupons.

The boxplot[Figure 1.1] further illustrates this trend, showing an upward shift in spending among coupon users, with some making significantly larger purchases, likely leveraging coupons for high-value transactions. In contrast, non-users demonstrate more consistent spending patterns, suggesting their purchasing behaviour is less influenced by promotional discounts. These insights suggest focusing on high spenders to drive coupon usage and refine marketing strategies.
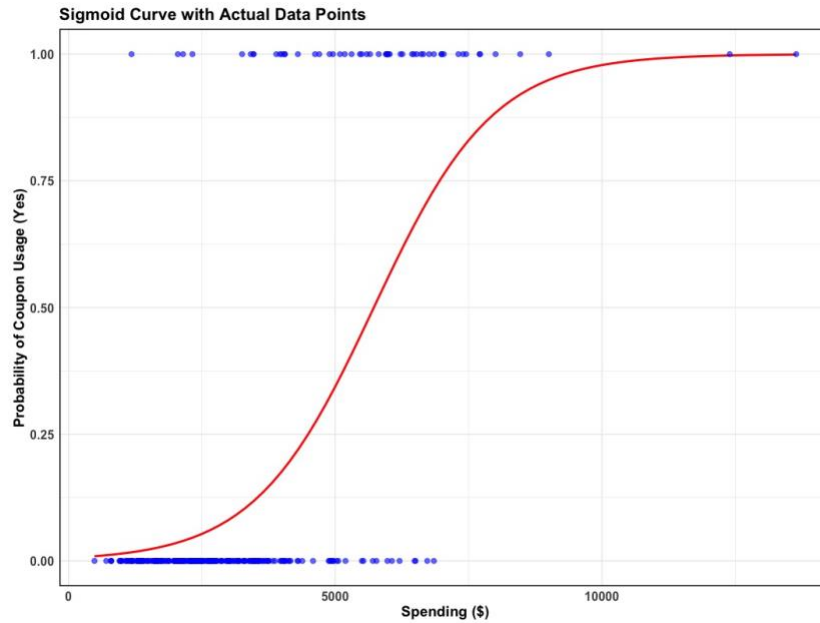
Figure 1.2 *Sigmoid Curve Representing the Probability of Coupon Redemption Based on Spending*

$$\text{Formula}: \hat{p} = \left(1 + e^{-(-5.11 + 0.000892 \cdot \mathit{Spending})}\right)^{-1}$$

Term	Estimate	p-value
Intercept	-5.11	2.00e-44 (<0.001)
Spending	0.000892	2.13e-29 (<0.001)

Table 1.2 *Logistic Regression Coefficients for Coupon Redemption*

The logistic regression model was designed to predict coupon redemption likelihood based on spending behaviour, using a stratified 70% training and 30% testing split to preserve the proportion of coupon users and non-users. As shown in[Figure. 1.2], the sigmoid curve indicates that at lower spending levels(below $2,000), the probability of coupon redemption remains low. However, as spending exceeds $5,000, the likelihood increases sharply, reaching near certainty at $10,000 and above. The spending coefficient(0.000892)[Table 1.2] suggests that each additional unit of spending increases the log-odds of coupon usage, corresponding to an odds ratio of exp(0.000892) ≈ 1.00089, indicating a slight increase in the probability of coupon redemption. A p-value of 2.13e-29, well below 0.001, confirms that spending is a statistically significant predictor of coupon adoption.
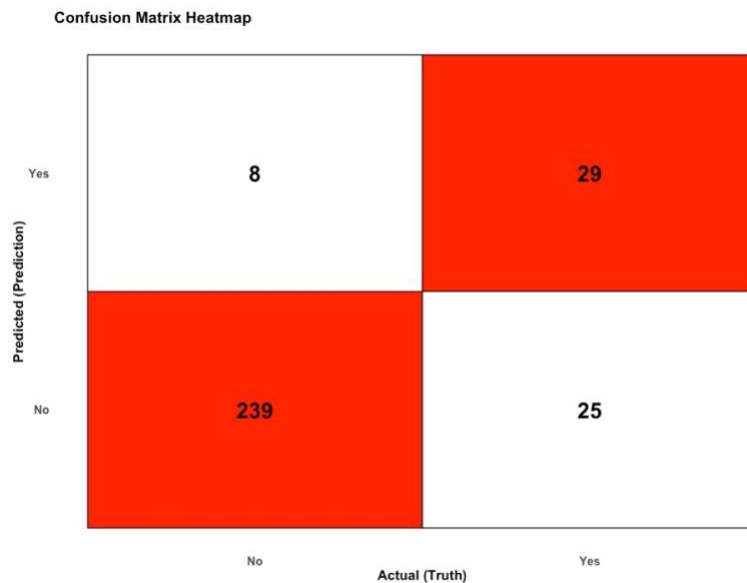
Figure 1.3 *Visual Representation of the Confusion Matrix*

ROC Curve for Coupon Redemption Model
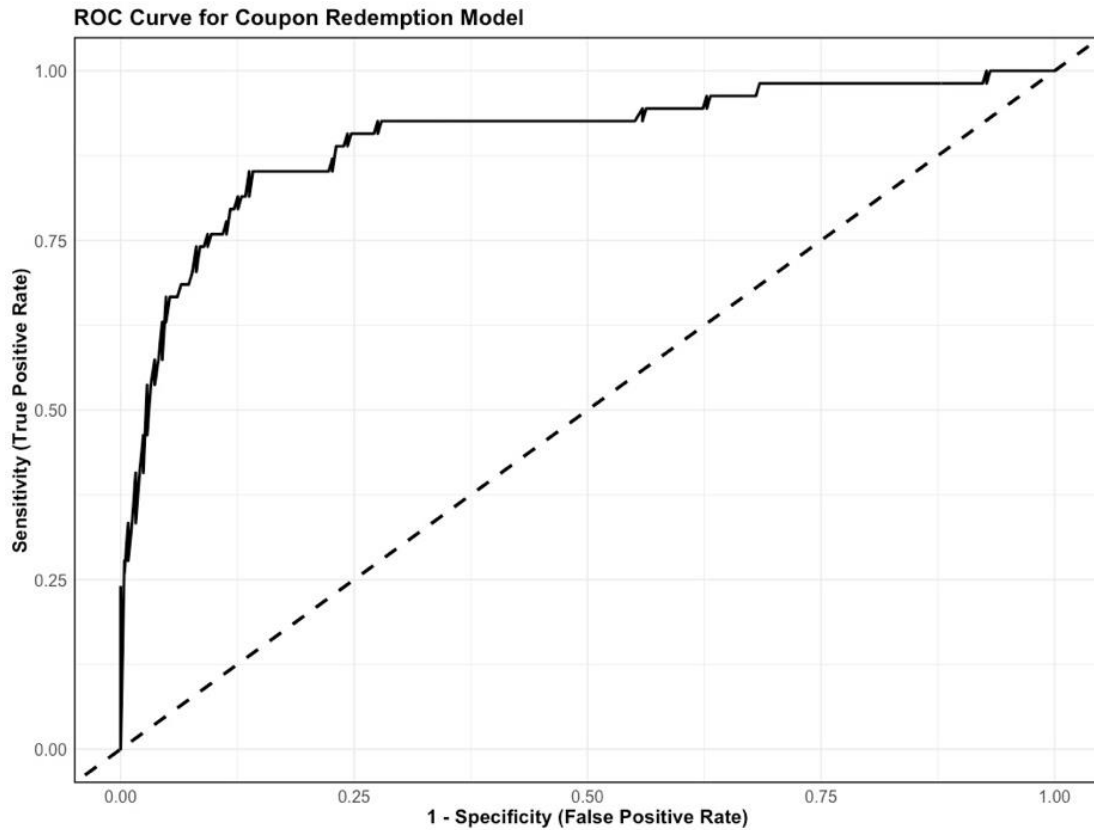
Figure 1.4 Model Performance Assessment Using ROC Curve

Metric	Accuracy	Sensitivity	Specificity	Precision	Area Under Curve (AUC)
Estimate	$0.890 \approx$ **89.0%**	$0.968 \approx$ **96.8%**	$0.537 \approx$ **53.7%**	$0.905 \approx$ **90.5%**	0.90

Table 1.3 Classification Metrics for the Logistic Regression Model

The model performance was evaluated using key classification metrics[Table 1.3]. The accuracy of 89.0% indicates that the model correctly classifies coupon users and non-users in most cases. The high sensitivity(96.8%) shows the model effectively identifies coupon users**(True Positives)**, with minimal misclassification as non-users**(False Negatives)**. This aligns with the confusion matrix (Figure 1.4), where 29 actual coupon users were correctly classified as users, and 8 non-users were misclassified as coupon users**(False Positives)**. However, the specificity (53.7%) is relatively low, indicating that the model struggles to correctly identify non-users, misclassifying 25 coupon users as non-users**(False Negatives)**. The high precision(90.5%) means that when the model predicts a customer will redeem a coupon, it is correct most of the time. Additionally, the AUC-ROC score of 0.90[Table 1.3] and[Figure 1.4] confirms strong discriminatory power, suggesting the model effectively differentiates between coupon users and non-users.

In summary, the logistic regression model demonstrates strong predictive capabilities, particularly in identifying coupon users. These insights highlight that spending is a strong predictor of coupon redemption, reinforcing the idea that higher-spending customers are more likely to use coupons.