Analysing the impact of spending on coupon usage using logistic regression

Author: Rohan Chawla

The dataset Salmon includes two key variables: spending, which represents the total amount a consumer spends, and coupon usage, which indicates whether a consumer used a discount coupon. This analysis's continuous and binary variables are spending and coupons, respectively.

The comparison of spending between coupon users and non-users suggests that those who use coupons generally tend to spend more. This indicates a positive relationship between expenditure and coupon usage. A logistic regression model(*Table 1.1)* was built to quantify the relationship between expenditure and coupon usage. The results show that the log odd of using a coupon increases for every additional dollar spent. The model (*Table 1.1)* estimates that each one-unit increase in spending results in a 0.095% increase in the likelihood of coupon usage. The statistical significance of this result, with a p-value of less than 0.001, confirms that spending has a meaningful effect on whether a consumer uses a coupon. The logistic regression curve (*Graph 1.2)* further illustrates this relationship, showing a gradual increase in coupon usage probability as spending increases.

A confusion matrix(*Graph 1.3)* was analyzed to assess the reliability of the model. The model(*Table 1.2)* achieved an accuracy of 87%, demonstrating strong predictive performance. Specificity, which measures the ability to classify non-coupon users correctly, was high at 95.9%, indicating that the model effectively identifies those who do not use coupons. However, sensitivity, which measures the ability to identify actual coupon users correctly, was moderate at 46.1%, suggesting that while the model performs well overall, it has some limitations in detecting all coupon users. Precision, which reflects how often a predicted coupon user used a coupon, was 70.7%.

The ROC curve(*Graph 1.4)* further assessed the model's ability to distinguish between coupon users and non-users. The area under the curve (AUC) was calculated at 0.881, suggesting that the model separated these groups. AUC values close to 1 indicate a strong predictive model, whereas values near 0.5 suggest random guessing. An AUC of 0.881 means that the model correctly ranks the coupon user higher 88.1% of the time.

In conclusion, our analysis indicates that increased spending positively influences coupon usage. The model effectively predicts non-users but has some limitations in identifying actual users. While it meets key assumptions, its sensitivity can be improved.

Table 1.1 Logistic Regression Model Coefficients

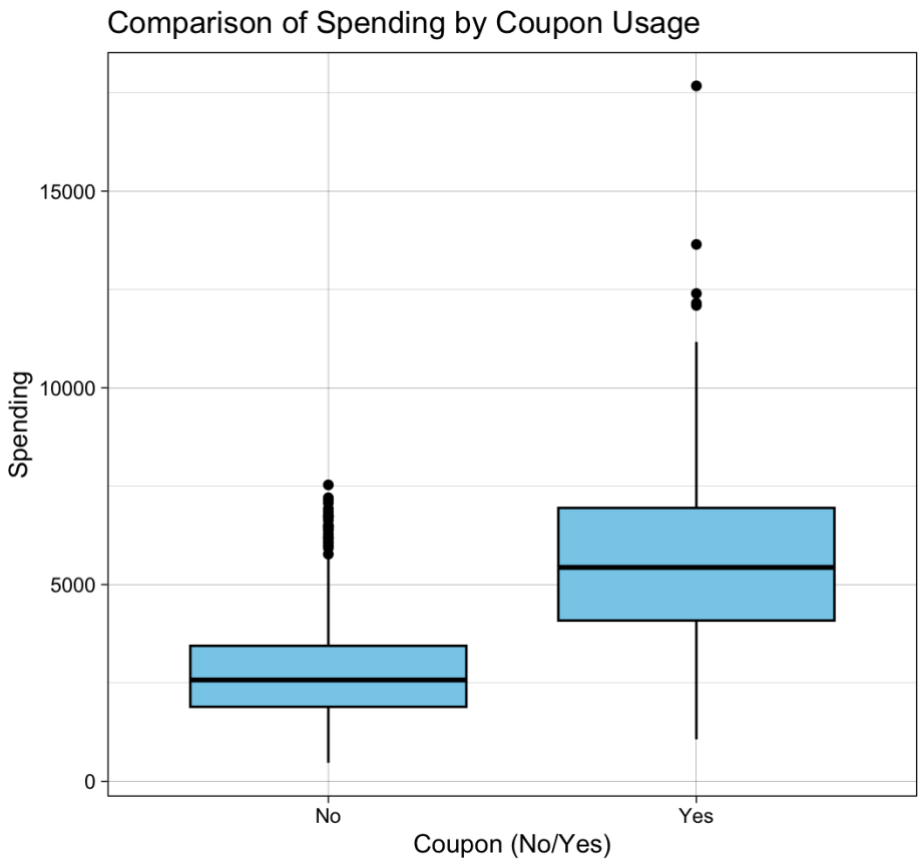
Coefficients	Estimate	P Value
Intercept	-5.333	<2e-16 (less than 0.005)
spending	0.0009511	<2e-16 (less than 0.005)

Table 1.2 Model Performance Metrics for Coupon Usage Prediction

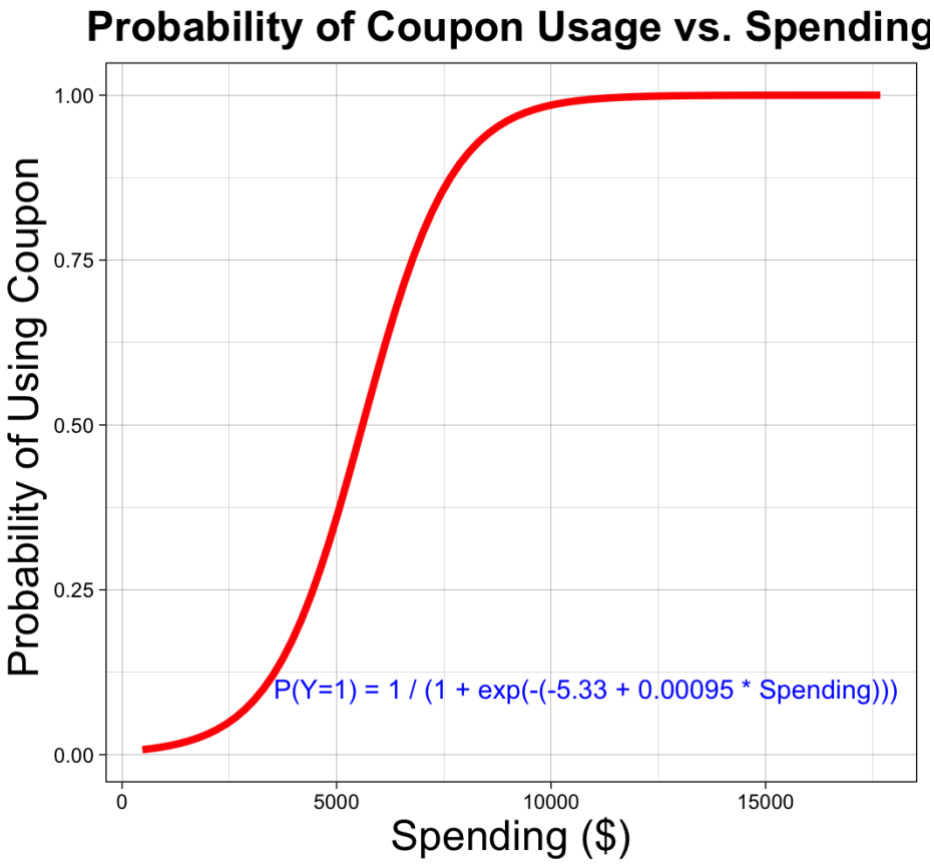
Metric	Accuracy	Sensitivity	Precision	Specificity	AUC (Area under curve)
Estimate	0.87≈87%	0.46≈46%	0.70≈70%	0.46≈46%	0.881

Graph 1.1 Boxplot showing Spending by Coupon usage

Reference:
https://chatgpt.com/#

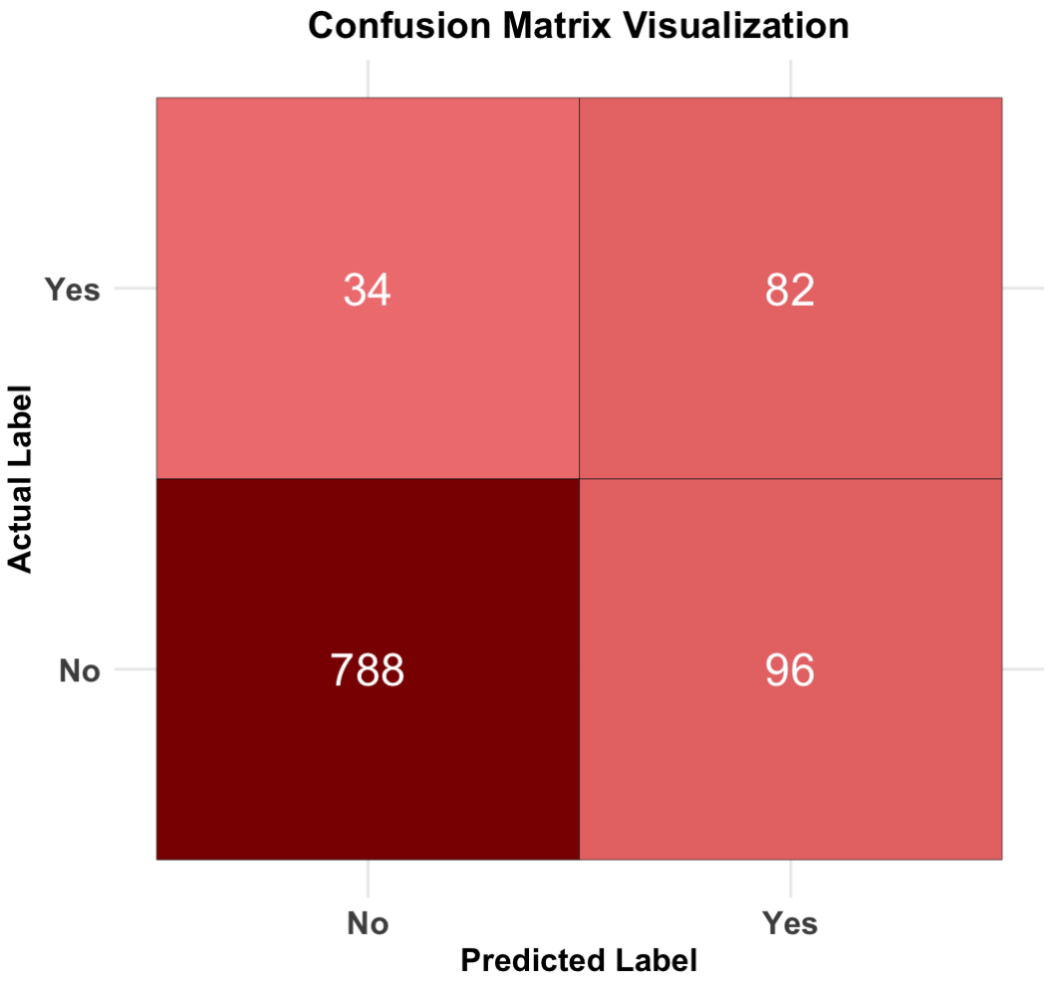


Graph 1.2 Sigmoid curve showing Probability of coupon usage with spending

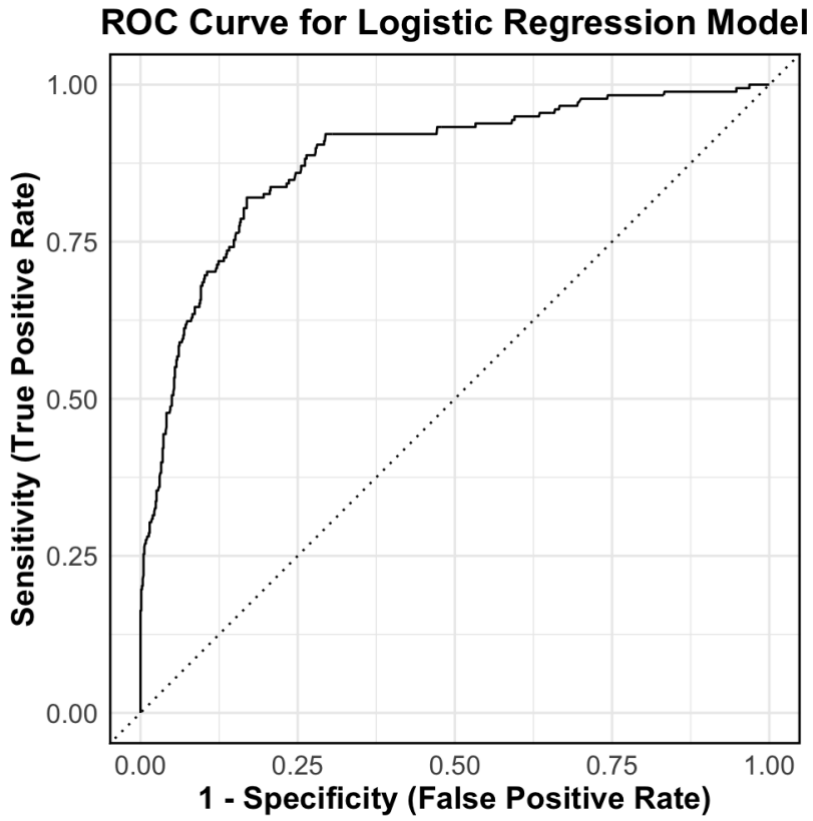


Reference:
https://chatgpt.com/#

Graph 1.3 Visual Representation of Confusion Matrix



Graph 1.4 Visual Representation of ROC Curve



Reference:
https://chatgpt.com/#