Logistic Regression Analysis of Shipping Data

Anjum Shams DS423

Introduction

The reliability of timely delivery in shipping operations is a critical aspect of customer satisfaction and operational efficiency. Motivated by personal experiences of package delays, this project aims to develop a predictive model to ascertain the likelihood of shipments arriving on time. Logistic Regression is employed to address this binary classification problem, utilizing a dataset comprising 10,999 shipment records. This report provides a comprehensive analysis of the data, model building, diagnostics, and evaluation processes undertaken to achieve a robust predictive model.

Dataset Description

The dataset used in this study contains 10,999 observations, with the dependent variable being Reached_on_time. This variable is binary, where a value of 1 indicates that the shipment did not reach on time, and a value of 0 indicates that it did. The dataset includes eleven independent variables: ID, Warehouse_block, Mode_of_Shipment, Customer_care_calls, Customer_rating, Cost_of_the_Product, Prior_purchases, Product_importance, Gender, Discount_offered, and Weight_in_gms.

To prepare the data for logistic regression analysis, the ID column was dropped as it serves as an identifier and does not contribute to the predictive modeling. Categorical variables (Warehouse_block, Mode_of_Shipment, Customer_rating, Product_importance, and Gender) were converted into dummy variables to facilitate their inclusion in the regression model. This transformation ensures that the categorical data is appropriately represented in the logistic regression framework.

Initial Data Analysis

A preliminary analysis of the frequency table revealed a slight imbalance in the dataset, with a higher proportion of shipments arriving on time compared to those delayed. Descriptive statistics highlighted some key insights: the median number of customer care calls was notably high, suggesting a correlation between delayed shipments and customer dissatisfaction. Additionally, the median discount offered was found to be less than 10%, indicating that discounts are generally modest.

Box plots were generated to visualize the distribution of shipment times across different modes of shipment and levels of product importance. The analysis showed minimal differences in delivery times among the various shipment modes (Flight, Road, Ship), with Flight having a slightly higher median, possibly due to differences in sample sizes. Similarly, the product importance categories (High, Medium, Low) exhibited minimal differences, though products of high importance had a slightly higher median delivery time.

Full Model Analysis

The initial logistic regression model included all variables, with the goal of maximizing the R-Square value while minimizing the AIC and SC values. The standardized estimates identified several key predictors: d_prod_imp_high, Discount_offered, Weight_in_gms, Customer_care_calls, and Prior_purchases.

The full logistic regression model to predict probability of reached on time p=Pr(reached_on_time=1) is fitted using PROC LOGISTIC:
log(p/1-p) = 1.75 - 0.04 d_WH_block_A +0.04 d_WH_block_B + 0.008 d_WH_block_C + 0.01 d_WH_block_D + 0.01d_MS_Ship + 0.03 d_MS_Flight -0.1 Customer_care_calls - 0.08 d_rating_1 - 0.11 d_rating_2 + 0.05 d_rating_3-0.02 d_rating_4 - 0.002 Cost_of_the_Product - 0.08 Prior_purchases - 0.3 d_prod_imp_low - 0.3 d_prod_imp_medium +0.05 d_Gender + 0.11 Discount_offered - 0.0002 Weight_in_gms

Multicollinearity and Diagnostics

To ensure the reliability of the model, multicollinearity was checked using the correlation matrix of the independent variables. No multicollinearity issues were detected as all correlation values were below 0.9. Additionally, the deviance residuals plot was examined to identify potential outliers. There were no significant outliers detected, with residuals within the acceptable range. Influential points were assessed using the Dfbeta statistic, and several influential points were identified and removed, although this led to minimal improvement in the model's performance.

Model Refinement and Selection

The dataset was split into a 70:30 ratio for training and testing purposes. Two model selection methods were employed: stepwise selection and backward selection. The stepwise selection method identified seven significant predictors with an R-Square value of 0.23, while the backward selection method identified a slightly different set of predictors with an R-Square value of 0.2382.

The final model included six significant predictors and demonstrated an improved R-Square value of 0.2386. 23.86% of the variation in reached_on_time is explained by the model, the rest is unexplained. The final logistic regression equation is:

log(reached_on_time =1/reached_on_time=0) = 1.61 - 0.23 Customer_care_calls + 0.16 d_rating_3 - 0.08 Prior_purchases + 0.33 d_prod_imp_high + 0.11Discount_offered - 0.00028 Weight_in_gms

The logistic regression model indicates that an increase in customer care calls by one unit is associated with a 20.55% decrease in the likelihood of the shipment reaching on time, while an increase in customer rating 3 by one unit corresponds to a 17.35% increase in on-time delivery likelihood. Additionally, each prior purchase increment decreases the likelihood of timely shipment by 7.69%. Conversely, high product importance boosts the on-time delivery probability by 39.09%, and a 1% increase in the discount offered enhances it by 11.63%. However, for each gram increase in product weight, the likelihood of on-time delivery decreases by 0.028%.

Model Validation

The goodness of fit for the final model was tested using the likelihood ratio test. The null hypothesis (H0: βj=0) was tested against the alternative hypothesis (Ha: βj≠0). The likelihood ratio was 2071.4736 with a P-value < 0.0001, leading to the rejection of the null hypothesis. This result confirms that at least one predictor has a significant association with the dependent variable.

Predicted probabilities and confidence intervals were computed for scenarios with one and three customer care calls, and high and low product importance. The predicted probabilities and confidence intervals for shipments reaching on time vary based on customer care calls and product importance. If a shipment has one customer care call and high product importance, the predicted probability of timely arrival is 84.7%, within a confidence interval of 120.34% to 141.08%. With three customer care calls and high product importance, the probability drops to 77.63%, with a confidence interval of 103.4% to 129.3%. For shipments with one customer care call but low product importance, the predicted probability is 79.8%, within a 109.6% to 129.3% confidence interval. Finally, with three customer care calls and low product importance, the predicted probability is 71.2%, with a confidence interval of 91.6% to 113.8%.

Classification and Performance Metrics

To evaluate the performance of the model, a confusion matrix was generated. For a threshold of 0.65, the model's test performance metrics were as follows:

- Accuracy: 66%
- Sensitivity: 47%
- Precision: 93%
- Specificity: 95%.

Model Comparison

The comparison of the two model selection methods (stepwise and backward) revealed that the stepwise selection method performed slightly better in terms of training performance. The final model obtained through stepwise selection had a higher R-Square value (23.86), lower AIC

(8220.479), and SC (8269.030) compared to the backward selection model. Both models exhibited identical test performance metrics, demonstrating robustness and reliability.

Conclusion

The logistic regression model developed in this project provides a moderate level of predictive accuracy for determining whether a shipment will reach on time. Key predictors include customer care calls, product importance, and discount offered, among others. The model can be further refined with additional data and advanced techniques to improve its robustness and accuracy. The stepwise selection method was slightly better than the backward selection method, both in training and overall performance.

Future Work

Dummy variables should be created by setting a base zero to avoid a biased model. Setting a base category allows for meaningful interpretation of the regression coefficients. Each dummy variable's coefficient represents the effect of that category relative to the base category. This relative interpretation is crucial for understanding the impact of different levels of the categorical variable on the dependent variable.

Recommendations

To enhance the predictive power of the model, it is recommended to increase the dataset size and incorporate more diverse observations. Additional features, such as weather conditions, and distance traveled, could be explored to provide a more comprehensive analysis.

Overall, this study highlights the importance of statistical analysis in understanding and improving logistics operations, ultimately leading to better customer satisfaction and operational efficiency.