LDA/QDA Code

2024-04-19

# Load MASS Library

library(MASS)

# Load Data Sets

order\_products\_train <- read.csv(“path\_to\_order\_products\_train.csv”)

orders <- read.csv(“path\_to\_orders.csv”)

# Combine Data Sets

combined\_data <- merge(order\_products\_train, orders, by = “order\_id”)

# Create churn variable

combined\_datareordered == 0)

# Ensure the ‘churn’ variable is a factor

combined\_datachurn)

# prepare data by selecting predictors and response

predictors <- combined\_data[, c(“order\_hour\_of\_day”, “days\_since\_prior\_order”, “add\_to\_cart\_order”)]

response <- combined\_data$churn

# Run LDA and QDA

lda\_model <- lda(response ~ ., data = data.frame(response, predictors))

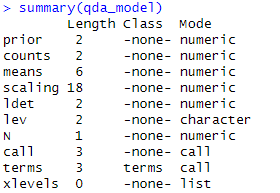
qda\_model <- qda(response ~ ., data = data.frame(response, predictors))

summary(lda\_model)

A screenshot of a computer

Description automatically generated

summary(qda\_model)



# Get Means and Scaling Coefficients

print(lda\_model

A black and white text

Description automatically generated

scaling)

A screenshot of a computer code

Description automatically generated

print(qda\_model$means)

A black and white text

Description automatically generated

# Get Predictions And Confusion Matrix

lda\_pred <- predict(lda\_model, newdata = predictors)

table(predicted = lda\_pred

A white background with black text

Description automatically generated

* Accuracy: Approximately 62.59%, indicating that the model correctly predicts churn for about 62.59% of the cases.
* Precision: Approximately 58.10%, indicating that when the model predicts churn, it is correct about 58.10% of the time.
* Recall: Approximately 24.43%, indicating that the model correctly identifies 24.43% of all actual churn cases.
* F1 Score: Approximately 34.40%, which is the harmonic mean of precision and recall and is a measure of the model's accuracy.

class, actual = response)

A white background with blue text

Description automatically generated

* Accuracy: Approximately 61.94%, indicating that the model correctly predicts churn for about 61.94% of the cases.
* Precision: Approximately 56.93%, indicating that when the model predicts churn, it is correct about 56.93% of the time.
* Recall: Approximately 21.30%, indicating that the model correctly identifies 21.30% of all actual churn cases.
* F1 Score: Approximately 31.01%, which is the harmonic mean of precision and recall and is a measure of the model's accuracy.

# Also Can Take Posterior Probabilities

lda\_pred

posterior

MAIN TAKEAWAYS FROM THIS MODEL:

* Both LDA and QDA models have moderate overall accuracy, with LDA slightly outperforming QDA.
* The precision of both models is moderate, indicating a reasonable level of reliability when they predict churn.
* The recall is low for both models, suggesting they miss a significant number of true churn instances.
* The F1 scores are relatively low, reflecting a need for improved balance between precision and recall.
* The most influential predictor in the LDA model is days\_since\_prior\_order, as it has the highest coefficient.
* The time of the day when the order is made (order\_hour\_of\_day) and the order in which items are added to the cart (add\_to\_cart\_order) are less influential but still significant predictors in the LDA model.
* Given the large dataset, the absolute number of misclassified instances is substantial, which can be impactful when considering business decisions based on the model’s predictions.
* The analysis indicates that while useful, LDA and QDA may not fully capture the complexity of the factors that contribute to churn, suggesting that further model refinement or alternative approaches may be necessary.