Boosting

2024-04-20

# Load necessary libraries

library(gbm)

# Load Data Sets

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

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

# Merge datasets on order\_id

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

# Convert ‘churn’ to a factor

combined\_datachurn)

# Convert ‘churn’ from factor to numeric (0 and 1)

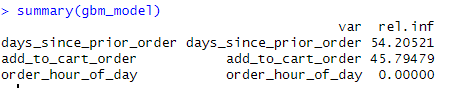
combined\_datachurn) - 1

# Fit the gradient boosting model

gbm\_model <- gbm(churn\_numeric ~ add\_to\_cart\_order + order\_hour\_of\_day + days\_since\_prior\_order, data = combined\_data, # make sure to use the updated dataset distribution = “bernoulli”, n.trees = 500, interaction.depth = 1, shrinkage = 0.01, cv.folds = 5, n.minobsinnode = 10, verbose = TRUE)

# Summary of the model to check variable importance

summary(gbm\_model)



A blue rectangular object with numbers

Description automatically generated

MAIN FINDINGS:

* Days Since Last Order: This variable is the most important predictor of churn, with a relative influence of approximately 54.2%. Customers are more likely to churn the longer it has been since their last order.
* Add to Cart Order: The add\_to\_cart\_order is also a significant predictor, with about 45.8% relative influence. The position in which an item is added to the shopping cart influences the likelihood of churn.
* Order Hour of Day: This variable has zero relative influence in the model, indicating it does not significantly affect churn prediction in the presence of the other variables.
* First Model to say order hour of day has absolutely no impact on churn