Classification Trees And Random Foresting

2024-04-19

# Load necessary libraries

library(rpart) library(ranger)

# 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)

# Select only the predictors and the response for the model

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

# Classification Tree with rpart

tree\_model <- rpart(churn ~ add\_to\_cart\_order + order\_hour\_of\_day + days\_since\_prior\_order, data = model\_data, method = “class”)

# Plot the tree

plot(tree\_model) text(tree\_model, use.n = TRUE)

A diagram of a diagram

Description automatically generated with medium confidence

# Load the ranger package

library(ranger)

# Fit the model with ranger

ranger\_rf\_model <- ranger( formula = churn ~ add\_to\_cart\_order + order\_hour\_of\_day + days\_since\_prior\_order, data = model\_data, num.trees = 500, write.forest = TRUE, # To save the forest for predictions importance = ‘impurity’ # To get variable importance )

# View the variable importance

ranger\_rf\_model$variable.importance

A close-up of a number

Description automatically generated

MAIN FINDINGS

* Days Since Prior Order: This predictor appears to be the most significant for churn prediction in both the classification tree and random forest model, suggesting that the longer the time since a customer's last order, the more likely they are to churn.
* Add to Cart Order: The second split in the classification tree is based on add\_to\_cart\_order. The random forest model also indicates it as an important feature, meaning the position in which an item is added to the cart may influence the likelihood of reorder.
* Random Forest Importance: The random forest has identified days\_since\_prior\_order as the most important predictor, followed by add\_to\_cart\_order, and then order\_hour\_of\_day, indicating that time since the last order and cart addition sequence are more influential than the hour of the day.
* Model Complexity: The classification tree provides a simple and interpretable model with two splits, which can be easily understood and acted upon. However, it may not capture the full complexity of the factors that influence churn.
* Random Forest Robustness: Random forests create a more robust and typically more accurate model by combining the predictions of many trees, but they are less interpretable than a single decision tree.
* Additional Analysis: While the classification tree and random forest provide initial insights, you could perform further analysis to validate these findings, such as cross-validation to assess model stability and generalizability. Alternate Models: Explore other machine learning models and compare their performance to ensure the best model is chosen for deployment.