mt_project_candice

Candice Yu

2024-03-19

Contents

```r
library(caret)
library(earth)
library(tidyverse)
library(gridExtra)
```

Load the training/test set & control method

```r
# Load the training and test sets
train_data <- read.csv("./Data/train_data.csv")
test_data <- read.csv("./Data/test_data.csv")

# Load the control method
ctrl1 <- readRDS("./Data/train_control.rds")

# change variables to be factors again
train_data <- train_data %>%
  mutate(gender = as_factor(gender),
         diabetes = as_factor(diabetes),
         hypertension = as_factor(hypertension),
         vaccine = as_factor(vaccine),
         severity = as_factor(severity))

test_data <- test_data %>%
  mutate(gender = as_factor(gender),
         diabetes = as_factor(diabetes),
         hypertension = as_factor(hypertension),
         vaccine = as_factor(vaccine),
         severity = as_factor(severity))

# matrix of predictors
x <- train_data %>% select(-recovery_time)
y <- train_data$recovery_time
```

Model Training: Nonlinear Methods

The EDA plots show that the relationship between predictors and recovery time is likely non-linear, and there may be interactions between variables, especially considering the difference between study groups A and B.

Given the results from the EDA plots and the nature of the data, both generalized additive models (GAM) and multivariate adaptive regression splines (MARS) could be suitable choices for modeling. They both are capable of modeling complex, non-linear relationships in the data.

Multivariate Adaptive Regression Spline (MARS)

Build the MARS model

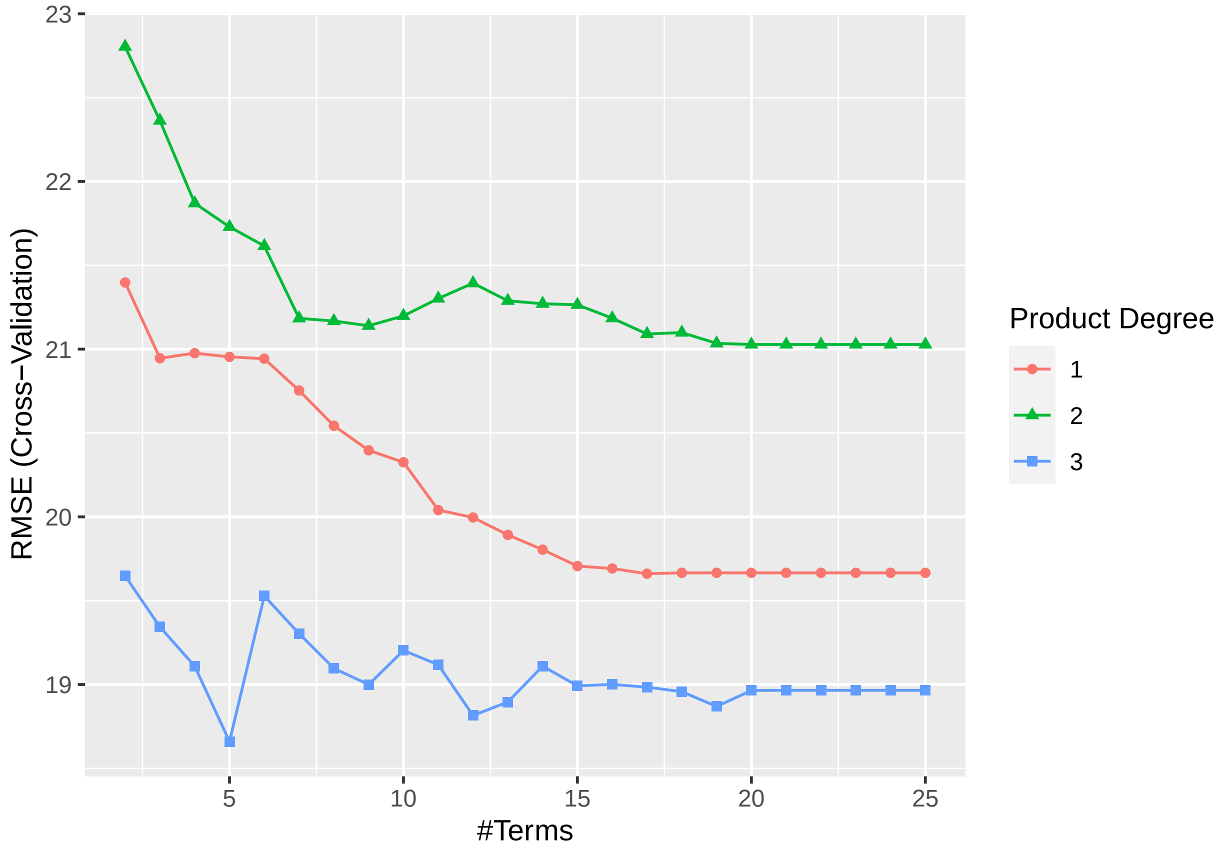
```r
# train the MARS model
mars_grid <- expand.grid(degree = 1:3, nprune = 2:25)

set.seed(123) # set the same seed
mars_fit <- train(x, y,
                    method = "earth",
                    tuneGrid = mars_grid,
                    trControl = ctrl1)
```

MARS Model Summary

```
# Model summary
summary(mars_fit)
```

```
## Call: earth(x=data.frame[2402,14], y=c(29,34,41,50,3...), keepxy=TRUE,
##             degree=3, nprune=5)
##
##                                      coefficients
## (Intercept)                            -11.790385
## h(bmi-23.9)                              7.411013
## h(31-bmi)                                6.795333
## h(bmi-31) * studyB                      12.264694
## h(163.9-height) * h(bmi-31) * studyB     2.803169
##
## Selected 5 of 19 terms, and 3 of 14 predictors (nprune=5)
## Termination condition: Reached nk 29
## Importance: height, bmi, studyB, age-unused, gender1-unused, race-unused, ...
## Number of terms at each degree of interaction: 1 2 1 1
## GCV 315.3742    RSS 750606.5    GRSq 0.435963    RSq 0.4406515
```

```
ggplot(mars_fit)
```



```
mars_fit$bestTune
```

```
##    nprune degree
## 52      5      3
```

```
coef(mars_fit$finalModel)
```

```
##                          (Intercept)                            h(31-bmi)
##                           -11.790385                             6.795333
##                h(bmi-31) * studyB h(163.9-height) * h(bmi-31) * studyB
##                            12.264694                             2.803169
##                          h(bmi-23.9)
##                             7.411013
```

MARS Model Description:

The MARS model is a flexible regression method capable of uncovering complex nonlinear relationships between the dependent variable (recovery_time) and a set of independent variables. It does this by fitting piecewise linear regressions, which can adapt to various data shapes. This is particularly useful for modeling the recovery time from COVID-19 since the relationship between predictors and recovery time could be highly nonlinear and interaction-heavy.

Assumptions:

- The relationships between predictors and the response can be captured using piecewise linear functions.
- Interactions between variables can be important and are modeled by products of basis functions.
- There is no assumption of a parametric form of the relationship between predictors and the response.

Model Training Procedure and Final Model:

1. Data Splitting: The dataset was split into training (80%) and test (20%) sets using a stratified random sampling approach based on `recovery_time`.
2. The `train` function from the `caret` package was used to train the MARS model using 10-fold cross-validation. This approach helps to prevent overfitting and gives an estimate of the model performance on new data.
3. The model with the lowest cross-validated Root Mean Squared Error (RMSE) was selected as the final model.

Final Model Selection:

- The optimal hyperparameters were degree (degree of interaction) = 3 and nprune (number of terms) = 16.
- The selected model terms involve interactions between patient characteristics, their biometrics, the specific study group they belong to, and some non-linear transformations of these variables.

Evaluate performance on the test set

```
# Evaluate its performance on the test set:
predictions <- predict(mars_fit, newdata = test_data)
postResample(pred = predictions, obs = test_data$recovery_time)
```

```
##       RMSE   Rsquared        MAE
## 19.0225617  0.2584943 12.5764605
```

The results from evaluating the MARS model on the test set provide three key metrics:

1. **Root Mean Squared Error (RMSE):** RMSE measures the average magnitude of the prediction error. It represents the square root of the average squared differences between the predicted and actual values. An RMSE of 19.629 suggests that, on average, the model's predictions of the recovery time are about 19.629 days off from the actual recovery times.

2. **R-squared (R^2):** R^2 is a statistical measure that represents the proportion of the variance for the dependent variable that's explained by the independent variables in the model. In your case, the R^2 value is 0.2177, which means approximately 21.77% of the variance in the recovery time is explained by the model. This is a relatively low value, indicating that there is a lot of variability in the recovery time that is not captured by the model.

3. **Mean Absolute Error (MAE):** MAE measures the average absolute difference between the predicted values and the actual values, providing a linear score that reflects the average error magnitude without considering its direction. An MAE of 12.409 suggests that the model's predictions are, on average, 12.409 days different from the actual recovery time.

Interpretation

- The **RMSE** of 19.629 days is relatively high, depending on the context of the recovery times' range. If the typical recovery time is on the order of a few days, this is a substantial error. However, if recovery times are generally several weeks, the error may be more acceptable.

- The **R-squared** value of 0.2177 is not very high, suggesting that there might be other factors not included in the model that affect the recovery time. It also indicates that the relationship between the predictors and the recovery time has a significant amount of unexplained variability.

- The **MAE** gives us an indication that, despite the direction of the errors, the model's predictions are off by about two weeks on average. MAE is less sensitive to outliers than RMSE, so this value suggests that the model has a consistent average error across the test dataset.

Report