

EXP 1: Linear Regression with RFE

AIM

The aim of this experiment is to build a Linear Regression model using the `mtcars` dataset to predict a car's fuel efficiency (measured in miles per gallon or mpg) based on various characteristics such as weight, horsepower, number of cylinders, and transmission type.

The experiment includes data loading, cross-validation, scaling, Recursive Feature Elimination (RFE) for feature selection, model training, summary evaluation, and making predictions on new car data.

EXPERIMENTAL SETUP

A CSV file or R's built-in dataset `mtcars` is used, which contains the following columns:

- `mpg` – Miles per gallon (fuel efficiency)
- `cyl` – Number of cylinders
- `disp` – Engine displacement
- `hp` – Gross horsepower
- `drat` – Rear axle ratio
- `wt` – Weight of the car
- `qsec` – 1/4 mile time
- `vs` – Engine type
- `am` – Transmission type
- `gear` – Number of forward gears
- `carb` – Number of carburetors

LIBRARIES REQUIRED

- `caret` – for feature selection, model training
- `ggplot2` – for visualization

STEPS

1. Load necessary libraries
2. Load the `mtcars` dataset
3. Setup 10-fold cross-validation
4. Scale and normalize features (handled by `caret`)
5. Perform Recursive Feature Elimination (RFE)
6. Visualize RFE results
7. Identify best subset of predictors
8. Train the Linear Regression model
9. Summarize and evaluate the model
10. Make predictions on new data

CODE SNIPPET

```
# Load the dataset
data(mtcars)

# Define target and predictor variables
target <- "mpg"
predictors <- setdiff(names(mtcars), target)
```

```
# Set up 10-fold cross-validation
train_control <- trainControl(
  method = "cv",
  number = 10,
  verboseIter = TRUE,
  savePredictions = "final",
  summaryFunction = defaultSummary,
  selectionFunction = "best"
)
```

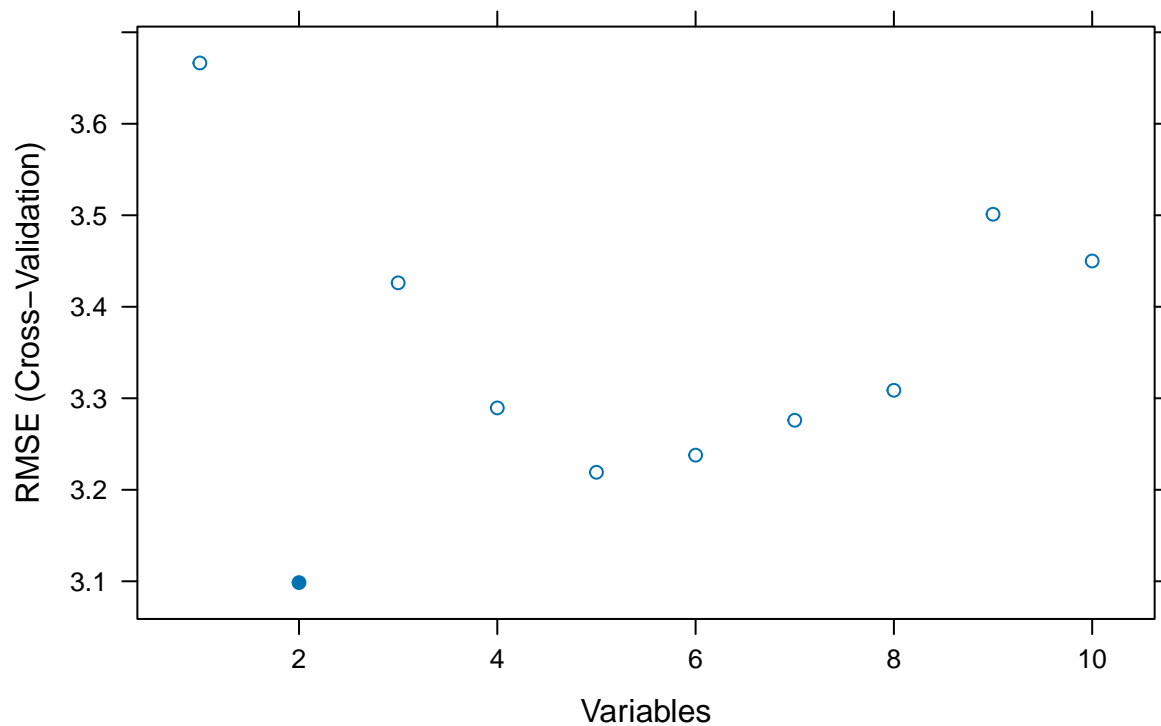
```
# RFE control using linear regression
rfe_control <- rfeControl(
  functions = lmFuncs,
  method = "cv",
  number = 10
)
```

```
# Run Recursive Feature Elimination
rfe_results <- rfe(
```

```
mtcars[, predictors],
mtcars[, target],
sizes = 1:length(predictors),
rfeControl = rfe_control
)
```

```
# Plot RFE results
plot(rfe_results, main = "RFE Results for Feature Selection in Linear Regression")
```

RFE Results for Feature Selection in Linear Regression



```
# Extract best features
best_features <- rfe_results$optVariables
print(best_features)
```

```
## [1] "wt" "am"
```

```
# Train final linear regression model with selected predictors
final_model <- lm(as.formula(paste(target, "~", paste(best_features, collapse = " + "))), data = mtcars)
```

```
# Display model summary
summary(final_model)
```

```
##
## Call:
```

```
## lm(formula = as.formula(paste(target, "~", paste(best_features,
##      collapse = " + "))), data = mtcars)
##
## Residuals:
##      Min        1Q    Median        3Q        Max
## -4.5295 -2.3619 -0.1317  1.4025  6.8782
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.32155    3.05464  12.218 5.84e-13 ***
## wt          -5.35281    0.78824  -6.791 1.87e-07 ***
## am          -0.02362    1.54565  -0.015  0.988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.098 on 29 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7358
## F-statistic: 44.17 on 2 and 29 DF,  p-value: 1.579e-09
```

```
# Create new data for prediction
```

```
new_cars_data <- data.frame(
  wt = c(2.5, 3, 3.5),
  hp = c(110, 140, 170),
  cyl = c(4, 6, 8),
  gear = c(4, 4, 5),
  drat = c(3.5, 3.2, 3.8),
  am = c(1, 0, 1),
  vs = c(0, 1, 0)
)
```

```
# Predict mpg for new data
```

```
predictions <- predict(final_model, newdata = new_cars_data)
print(predictions)
```

```
##           1           2           3
## 23.91591 21.26312 18.56310
```

CONCLUSION

The RFE process effectively selected a subset of features most relevant for predicting mpg, improving model interpretability and reducing overfitting. This experiment demonstrates the use of feature selection techniques to enhance model performance in real-world regression problems.