AMS_580_Q1

Sriram Vivek

2025-02-19

```r
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```r
library(MASS)
library(leaps)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Q1. Please use the random seed 123 to divide the data into 75% training and 25% testing.

```r
ames_data <- read.csv("/Users/sriram/Desktop/AMS 580/Ames_Housing_Data.csv")


ames_data <- na.omit(ames_data)


set.seed(123)


index <- createDataPartition(ames_data$SalePrice, p = 0.75, list = FALSE)
train_data <- ames_data[index, ]
test_data <- ames_data[-index, ]
```

Q2. Please find the best model using the stepwise variable selection method (based on the BIC criterion) using the training data.Please (a) display the coefficients of the fitted model; (b) make prediction on the testing data, and report the RMSE and the Coefficient of Determination R.

```
full_model <- lm(SalePrice ~ ., data = train_data)


step_model <- stepAIC(full_model, direction = "both", k = log(nrow(train_data)), trace = FALSE)


cat("Stepwise Model Coefficients:\n")
```

```
## Stepwise Model Coefficients:
```

```
print(coef(step_model))
```

```
##   (Intercept)       LotArea   OverallQual   OverallCond     YearBuilt
## -1.017464e+06  7.339645e-01  1.667461e+04  5.943478e+03  4.812917e+02
##      X1stFlrSF     X2ndFlrSF  BedroomAbvGr  KitchenAbvGr  TotRmsAbvGrd
##  1.013470e+02  6.477953e+01 -1.488137e+04 -3.342536e+04  4.361293e+03
##    GarageArea
##  3.559750e+01
```

```
step_predictions <- predict(step_model, newdata = test_data)


step_rmse <- sqrt(mean((test_data$SalePrice - step_predictions)^2))
step_r_squared <- cor(test_data$SalePrice, step_predictions)^2

cat("\nStepwise Model Performance:\n")
```

```
##
## Stepwise Model Performance:
```

```
cat("RMSE:", step_rmse, "\n")
```

```
## RMSE: 48727.68
```

```
cat("R^2:", step_r_squared, "\n")
```

```
## R^2: 0.6601152
```

Q3. Please find the best model using the best subset variable selection method (based on the SSE criterion) using the training data. Please (a) display the coefficients of the fitted model; (b) make prediction on the testing data, and report the RMSE and the Coefficient of Determination R2

```
subset_model <- regsubsets(SalePrice ~ ., data = train_data, nvmax = 20)

best_subset_index <- which.min(summary(subset_model)$bic)
best_subset_vars <- names(coef(subset_model, id = best_subset_index))[-1]  # Exclude intercept
```

```r
best_subset_formula <- as.formula(paste("SalePrice ~", paste(best_subset_vars, collapse = " + ")))
best_subset_model <- lm(best_subset_formula, data = train_data)

cat("\nBest Subset Model Coefficients:\n")
```

```
##
## Best Subset Model Coefficients:
```

```r
print(coef(best_subset_model))
```

```
##   (Intercept)       LotArea   OverallQual   OverallCond     YearBuilt
## -1.017464e+06  7.339645e-01  1.667461e+04  5.943478e+03  4.812917e+02
##     X1stFlrSF     X2ndFlrSF  BedroomAbvGr  KitchenAbvGr  TotRmsAbvGrd
##  1.013470e+02  6.477953e+01 -1.488137e+04 -3.342536e+04  4.361293e+03
##    GarageArea
##  3.559750e+01
```

```r
best_subset_predictions <- predict(best_subset_model, newdata = test_data)

best_subset_rmse <- sqrt(mean((test_data$SalePrice - best_subset_predictions)^2))
best_subset_r_squared <- cor(test_data$SalePrice, best_subset_predictions)^2

cat("\nBest Subset Model Performance:\n")
```

```
##
## Best Subset Model Performance:
```

```r
cat("RMSE:", best_subset_rmse, "\n")
```

```
## RMSE: 48727.68
```

```r
cat("R^2:", best_subset_r_squared, "\n")
```

```
## R^2: 0.6601152
```

Q4. Which model selection method among the 2 we have used above is the best? (a)Please compare the BIC of these models using the training data, as well as display these two models so we can see the parameter estimators and model goodness of fit measures. (b) Furthermore, please compare the RMSE and R2 of these models using the test data. (c) Please discuss any modifications you can do to further improve your model(s).

```r
step_bic <- BIC(step_model)
best_subset_bic <- BIC(best_subset_model)

cat("\nModel Comparison (BIC):\n")
```

```
##
## Model Comparison (BIC):
```

```r
cat("Stepwise Model BIC:", step_bic, "\n")
```

```
## Stepwise Model BIC: 26016.31
```

```r
cat("Best Subset Model BIC:", best_subset_bic, "\n")
```

```
## Best Subset Model BIC: 26016.31
```

```r
cat("\nModel Comparison (Test Data):\n")
```

```
##
## Model Comparison (Test Data):
```

```r
cat("Stepwise Model RMSE:", step_rmse, "\n")
```

```
## Stepwise Model RMSE: 48727.68
```

```r
cat("Best Subset Model RMSE:", best_subset_rmse, "\n")
```

```
## Best Subset Model RMSE: 48727.68
```

```r
cat("Stepwise Model R^2:", step_r_squared, "\n")
```

```
## Stepwise Model R^2: 0.6601152
```

```r
cat("Best Subset Model R^2:", best_subset_r_squared, "\n")
```

```
## Best Subset Model R^2: 0.6601152
```

```r
cat("\nPotential Improvements:\n")
```

```
##
## Potential Improvements:
```

```r
cat("1. Interaction terms or polynomial terms could be added to capture non-linear relationships.\n")
```

```
## 1. Interaction terms or polynomial terms could be added to capture non-linear relationships.
```

```r
cat("2. Regularization techniques can be used, like Ridge or Lasso regression to handle multicollineari
```

```
## 2. Regularization techniques can be used, like Ridge or Lasso regression to handle multicollinearity
```

```r
cat("3. Feature engineering, such as log transformations can be performed for skewed predictors.\n")
```

```
## 3. Feature engineering, such as log transformations can be performed for skewed predictors.
```

```r
cat("4. More advanced models like Random Forest or Gradient Boosting could be explored for better predic
```

```
## 4. More advanced models like Random Forest or Gradient Boosting could be explored for better predicti
```