WhiteWineLogisticRegression

2024-05-06

Logistic Regression on White Wine

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
r = getOption("repos")
r["CRAN"] = "http://cran.us.r-project.org"
options(repos = r)
install.packages('Metrics')
```

```
##
## The downloaded binary packages are in
##  /var/folders/5z/c3rg13j54bdfj4dsnvyvgsyh0000gn/T//RtmpfJqJIE/downloaded_packages
```

```
library(Metrics)
```

```
##
## Attaching package: 'Metrics'
```

```
## The following objects are masked from 'package:caret':
##
##     precision, recall
```

```
data <- read.csv("winequality-white.csv", header = TRUE, sep = ";")

# Binarize the quality variable for logistic regression
data$quality <- factor(ifelse(data$quality >= 7, "High", "Low"))

# LOOCV
fitControl <- trainControl(method = "LOOCV", classProbs = TRUE) # Ensure class probabilities can be cal
model <- train(quality ~ ., data = data, method = "glm", family = "binomial", trControl = fitControl)

# Prediction
predictions <- predict(model, newdata = data, type = "prob")[,"High"]

# Test MSE
test_mse <- mse(as.numeric(data$quality == "High"), predictions)

# Adjust R squared in terms of classification
predicted_classes <- ifelse(predictions > 0.5, "High", "Low")
R2 <- R2(as.numeric(predicted_classes == "High"), as.numeric(data$quality == "High"), form = "adj")

# Relevant metrics for logistic regression
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
## The following object is masked from 'package:Metrics':
##
##     auc
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
roc_result <- roc(response = as.numeric(data$quality == "High"), predictor = predictions)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
auc_value <- auc(roc_result)
conf_matrix <- confusionMatrix(predict(model, newdata = data), data$quality)

list(MSE = test_mse, Adjusted_R2 = R2, AUC = auc_value, Accuracy = conf_matrix$overall['Accuracy'], Sens
```

```
## $MSE
## [1] 0.1350475
##
## $Adjusted_R2
## NULL
##
## $AUC
## Area under the curve: 0.7938
##
## $Accuracy
##  Accuracy
## 0.8023683
##
## $Sensitivity
## Sensitivity
##   0.2801887
##
## $Specificity
## Specificity
##   0.9465868
```

Logistic Regression on Red Wine

```
library(caret)
library(pROC)

data <- read.csv("winequality-red.csv", header = TRUE, sep = ";")

# Binarize the quality variable for logistic regression
data$quality <- factor(ifelse(data$quality >= 7, "High", "Low"))

# LOOCV for logistic regression
fitControl <- trainControl(method = "LOOCV", classProbs = TRUE, summaryFunction = twoClassSummary) # Ad
```

```r
model <- train(quality ~ ., data = data, method = "glm", family = "binomial", trControl = fitControl, m

# Prediction
predictions <- predict(model, newdata = data, type = "prob")[,"High"]

# Test MSE
test_mse <- mse(as.numeric(data$quality == "High"), predictions)

# Other relevant metrics
roc_result <- roc(response = as.numeric(data$quality == "High"), predictor = predictions)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```r
auc_value <- auc(roc_result)
conf_matrix <- confusionMatrix(predict(model, newdata = data), data$quality)

# Output
list(MSE = test_mse, AUC = auc_value, Accuracy = conf_matrix$overall['Accuracy'], Sensitivity = conf_ma
```

```
## $MSE
## [1] 0.08422009
##
## $AUC
## Area under the curve: 0.8822
##
## $Accuracy
##  Accuracy
## 0.8843027
##
## $Sensitivity
## Sensitivity
##   0.3456221
##
## $Specificity
## Specificity
##   0.9688857
```

In this logistic regression model, we attempted to replicate similar logistic regression models we learned in class to predict the quality of white wine. However, in order to apply the model well, we needed to binarize the quality variable to fit a logistic regression model. The first aspect we tackle is binarizing the quality variable into High and Low. In order to perform effective comparative analysis with the other models, we decided to use LOOCV as standard cross validation measure. Since we are using the most optimal k-fold cross validation method, LOOCV, and binarizing the quality variable to fit the logistic regression, post consulation, we decided that it would be best to fit the model directly rather than add a layer of variable selection. Best subset selection and stepwise selection performed worse and had been too computationally intensive. Thus, to fit a simple logistic regression model and achieve the most optimal outcome metrics, we programmed the metrics that are most relevant to a logistic regression model, including test MSE, AUC, Accuracy, Sensitivity, and Specificity, which encompass AUC, ROC, and a Confusion Matrix. The same process is applied to red wine. Our combinations of the methods resulted in the most optimal outcome metrics.