Wines LDA QDA Classification

Isabella Chen

2025-04-30

```
library(MASS)
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```
library(ggplot2)
library(caret)

wine_data <- read.csv("/Users/isabellachen/Downloads/wine-quality-white-and-red.csv")

wine_data$quality_bin <- factor(ifelse(wine_data$quality >= 7, "high", "low"))

# Split into 70/30 train/test
set.seed(1)
train_idx   <- createDataPartition(wine_data$quality_bin, p = 0.7, list = FALSE)
train_data  <- wine_data[train_idx, ]
test_data   <- wine_data[-train_idx, ]

# All predictors except 'quality'
fmla <- as.formula("quality_bin ~ . - quality")

# 4. Fit LDA and QDA
lda_model <- lda(fmla, data = train_data)
qda_model <- qda(fmla, data = train_data)

# 5. Predict on test set
lda_pred <- predict(lda_model, test_data)
qda_pred <- predict(qda_model, test_data)
```

```
# Evaluate LDA model
lda_conf <- confusionMatrix(lda_pred$class, test_data$quality_bin)
cat("LDA Model Performance:\n")
```

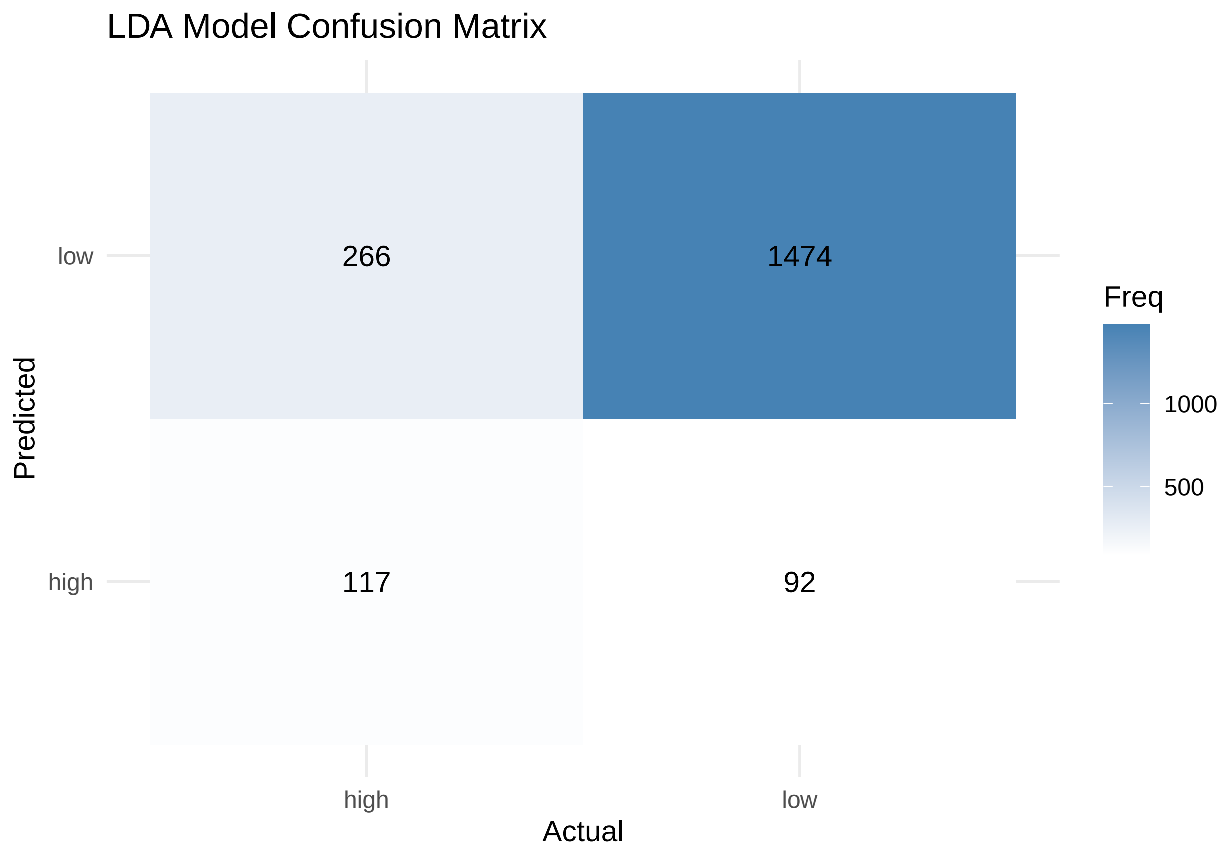
```
## LDA Model Performance:
```

```
print(lda_conf)
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction high  low
##       high  117   92
##       low   266 1474
##                                           
##                Accuracy : 0.8163          
##                  95% CI : (0.7984, 0.8333)
##     No Information Rate : 0.8035          
##     P-Value [Acc > NIR] : 0.08041         
##                                           
##                   Kappa : 0.2978          
##                                           
##  Mcnemar's Test P-Value : < 2e-16         
##                                           
##             Sensitivity : 0.30548         
##             Specificity : 0.94125         
##          Pos Pred Value : 0.55981         
##          Neg Pred Value : 0.84713         
##              Prevalence : 0.19651         
##          Detection Rate : 0.06003         
##    Detection Prevalence : 0.10723         
##       Balanced Accuracy : 0.62337         
##                                           
##        'Positive' Class : high            
## 
```

```r
lda_cm_df <- as.data.frame(lda_conf$table)
colnames(lda_cm_df) <- c("Predicted", "Actual", "Freq")

ggplot(lda_cm_df, aes(x = Actual, y = Predicted, fill = Freq)) +
  geom_tile() +
  geom_text(aes(label = Freq), color = "black") +
  scale_fill_gradient(low = "white", high = "steelblue") +
  labs(title = "LDA Model Confusion Matrix") +
  theme_minimal()
```



```r
# Evaluate QDA model
qda_conf <- confusionMatrix(qda_pred$class, test_data$quality_bin)
print("QDA Model Performance:")
```

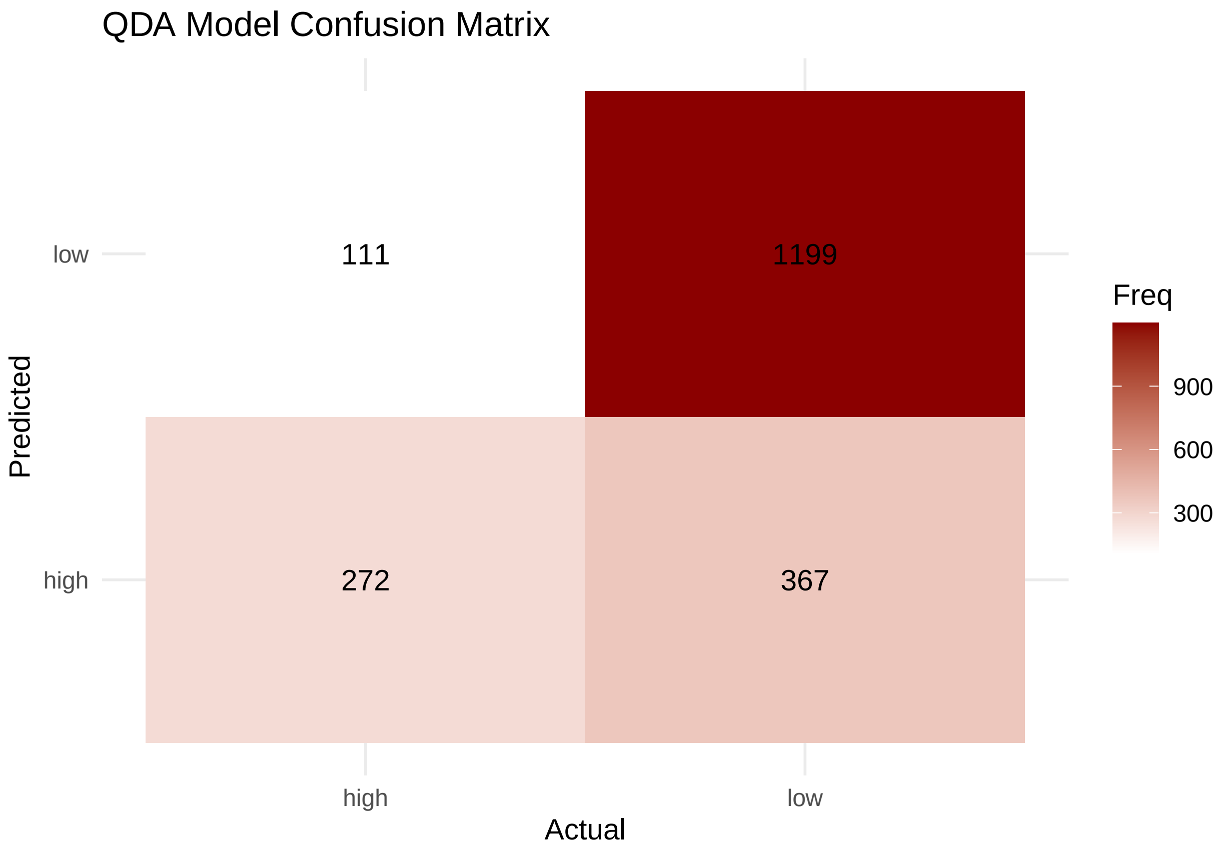
```
## [1] "QDA Model Performance:"
```

```r
print(qda_conf)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction high  low
##       high  272  367
##       low   111 1199
##
##                Accuracy : 0.7547
##                  95% CI : (0.735, 0.7737)
##     No Information Rate : 0.8035
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3799
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7102
##             Specificity : 0.7656
##          Pos Pred Value : 0.4257
##          Neg Pred Value : 0.9153
##              Prevalence : 0.1965
##          Detection Rate : 0.1396
##    Detection Prevalence : 0.3279
##       Balanced Accuracy : 0.7379
##
##        'Positive' Class : high
##
```

```
qda_cm_df <- as.data.frame(qda_conf$table)
colnames(qda_cm_df) <- c("Predicted", "Actual", "Freq")

ggplot(qda_cm_df, aes(x = Actual, y = Predicted, fill = Freq)) +
  geom_tile() +
  geom_text(aes(label = Freq), color = "black") +
  scale_fill_gradient(low = "white", high = "darkred") +
  labs(title = "QDA Model Confusion Matrix") +
  theme_minimal()
```



```
# Compare models
models <- c("LDA", "QDA")
accuracies <- c(
  lda_conf$overall["Accuracy"],
  qda_conf$overall["Accuracy"]
)
comparison <- data.frame(Model = models, Accuracy = accuracies)
print(comparison)
```

```
##   Model  Accuracy
## 1   LDA 0.8163161
## 2   QDA 0.7547460
```

```
set.seed(1)
# Cross validation with 10 fold
ctrl <- trainControl(method = "cv", number = 10)

lda_cv <- train(
  fmla,
  data      = train_data,
  method    = "lda",
  trControl = ctrl,
  preProcess= "nzv"
)

qda_cv <- train(
  fmla,
  data      = train_data,
  method    = "qda",
  trControl = ctrl,
  preProcess= "nzv"
)

print(lda_cv)
```

```
## Linear Discriminant Analysis
##
## 4548 samples
##   13 predictor
##    2 classes: 'high', 'low'
##
## Pre-processing:  (None)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 4094, 4093, 4094, 4093, 4092, 4093, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.8181559  0.3107644
```

```
print(qda_cv)
```

```
## Quadratic Discriminant Analysis
##
## 4548 samples
##   13 predictor
##    2 classes: 'high', 'low'
##
## Pre-processing:  (None)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 4093, 4092, 4093, 4094, 4093, 4093, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.7634141  0.3960156
```