Data Taming Assignment 3

Dongju Ma

2024-08-10

Setup

```
library(tidyverse)
library(tidymodels)
library(car)
library(modelr)
```

Q1. Loading the data

```
# Your student number goes here
ysn = 1942340
# Calculate your student number modulo 3
filenum <- (ysn + 2) %% 3
filenum
```

```
## [1] 1
```

```
filename <- paste0("./data/merry_",filenum,".csv")
filename
```

```
## [1] "./data/merry_1.csv"
```

```
# Read the csv file
merry <- read_csv(filename)
merry
```

```
## # A tibble: 30,000 x 6
##    RHBMM Accuracyyy AGE    DRESS Home   Jail
##    <dbl>      <dbl> <chr>  <chr> <chr>  <chr>
##  1     1      0.791 middle green forest yes
##  2     1      0.505 middle red   city   no
##  3     0      0.581 middle black city   no
##  4     0      0.735 middle red   city   no
##  5     0      0.530 youth  green city   no
##  6     0      0.432 youth  black city   no
```

```
##  7     0      0.758 middle red   city   no
##  8     0      0.568 youth  red   city   no
##  9     1      0.514 middle black city   no
## 10     1      0.764 youth  black forest no
## # i 29,990 more rows
```

```r
# Output the dimensions
dim(merry)
```

```
## [1] 30000     6
```

Q2. Identifying data types

- RHBMM: Categorical Nominal
  This column contains only two types of values: 1 and 0, representing whether the archer is married or not. Since these values are not ordered, this data should be classified as Categorical Nominal
- Accuracy: Quantitative Continuous
  This column represents the proportion of successful hits out of all attempts, with values ranging continuously between 0 and 1. So it should be classified as Quantitative Continuous.
- AGE: Categorical Ordinal
  This column categorizes archers into three age groups, ordered sequentially from youth to middle age, and finally to senior. So it should be classified as Categorical Ordinal.
- DRESS: Categorical Nominal
  This column shows the colors of each archer's outfit. There are three colors and the colors are not ordered, so it should be classified as Categorical Nominal.
- Home: Categorical Nominal
  This column only contains two kinds of values which are forest and city. Between these values there are no special orders. So it should be classified as Categorical Nominal.
- Jail: Categorical Nominal
  This column contains yes and no these two types of values which represents the archer was in jail before or not. These values are not ordered either so it still should be Categorical Nominal.

Q3. Taming Data

```r
# Change the column names
merry <- rename(merry,  rhbmm = RHBMM, acc = Accuracyyy, age = AGE, dress = DRESS,
                home = Home, jail = Jail)
# Transform into factors
merry$rhbmm <- as.factor(merry$rhbmm)
merry$age <- as.factor(merry$age)
merry$dress <- as.factor(merry$dress)
merry$home <- as.factor(merry$home)
merry$jail <- ifelse(merry$jail == 'yes', TRUE, FALSE)
merry$jail <- as.logical(merry$jail)
# Show the final version and the dimensions
merry
```

```
## # A tibble: 30,000 x 6
##    rhbmm   acc age    dress home   jail
```

```
##    <fct> <dbl> <fct>  <fct> <fct>  <lgl>
##  1 1     0.791 middle green forest TRUE
##  2 1     0.505 middle red   city   FALSE
##  3 0     0.581 middle black city   FALSE
##  4 0     0.735 middle red   city   FALSE
##  5 0     0.530 youth  green city   FALSE
##  6 0     0.432 youth  black city   FALSE
##  7 0     0.758 middle red   city   FALSE
##  8 0     0.568 youth  red   city   FALSE
##  9 1     0.514 middle black city   FALSE
## 10 1     0.764 youth  black forest FALSE
## # i 29,990 more rows
```

```
dim(merry)
```

```
## [1] 30000     6
```

Q4. Set the training set and testing set

```
# Set seed
set.seed(1942340)
# Split the data set
merry_split <- initial_split(merry, prop = 2/3)
merry_train <- training(merry_split)
merry_test <- testing(merry_split)
# Show the results
merry_split
```

```
## <Training/Testing/Total>
## <20000/10000/30000>
```

```
merry_train
```

```
## # A tibble: 20,000 x 6
##    rhbmm   acc age    dress home   jail
##    <fct> <dbl> <fct>  <fct> <fct>  <lgl>
##  1 0     0.844 youth  green city   FALSE
##  2 1     0.697 middle black city   FALSE
##  3 1     0.713 middle red   city   FALSE
##  4 1     0.694 senior black city   TRUE
##  5 0     0.742 middle red   city   FALSE
##  6 1     0.755 youth  black forest FALSE
##  7 1     0.787 youth  green forest FALSE
##  8 0     0.514 youth  red   city   FALSE
##  9 0     0.675 middle black city   FALSE
## 10 1     0.823 middle green city   FALSE
## # i 19,990 more rows
```

```r
dim(merry_train)
```

```
## [1] 20000     6
```

```r
merry_test
```

```
## # A tibble: 10,000 x 6
##    rhbmm   acc age    dress home   jail
##    <fct> <dbl> <fct>  <fct> <fct>  <lgl>
##  1 0     0.530 youth  green city   FALSE
##  2 1     0.764 youth  black forest FALSE
##  3 1     0.697 senior black forest FALSE
##  4 1     0.690 youth  black city   FALSE
##  5 0     0.854 youth  red   city   FALSE
##  6 1     0.665 senior green city   FALSE
##  7 0     0.833 middle green city   FALSE
##  8 1     0.759 youth  black forest FALSE
##  9 1     0.639 middle red   city   FALSE
## 10 1     0.703 senior black city   FALSE
## # i 9,990 more rows
```

```r
dim(merry_test)
```

```
## [1] 10000     6
```

Q5. Build a logistic regression model

```r
classification_lr <- logistic_reg() %>%
  set_engine('glm')
# Fit a logistic regression model with each variable individually
lrfit <- classification_lr %>%
  fit(rhbmm ~ acc+age+dress+home+jail, data = merry_train)
summary(lrfit$fit)
```

```
##
## Call:
## stats::glm(formula = rhbmm ~ acc + age + dress + home + jail,
##     family = stats::binomial, data = data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.504286   0.094125 -15.982  < 2e-16 ***
## acc          1.673335   0.126557  13.222  < 2e-16 ***
## agesenior   -0.043566   0.043805  -0.995  0.31995
## ageyouth     0.002011   0.035471   0.057  0.95479
## dressgreen   0.993479   0.038959  25.501  < 2e-16 ***
## dressred    -0.463103   0.037222 -12.442  < 2e-16 ***
## homeforest   1.727748   0.038633  44.722  < 2e-16 ***
```

```
## jailTRUE     0.171036   0.044696   3.827  0.00013 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27400  on 19999  degrees of freedom
## Residual deviance: 23619  on 19992  degrees of freedom
## AIC: 23635
##
## Number of Fisher Scoring iterations: 4
```

Q6. Fit a matrix model

```
model_matrix(merry_train, ~dress+age)
```

```
## # A tibble: 20,000 x 5
##    `(Intercept)` dressgreen dressred agesenior ageyouth
##            <dbl>      <dbl>    <dbl>     <dbl>    <dbl>
##  1             1          1        0         0        1
##  2             1          0        0         0        0
##  3             1          0        1         0        0
##  4             1          0        0         1        0
##  5             1          0        1         0        0
##  6             1          0        0         0        1
##  7             1          1        0         0        1
##  8             1          0        1         0        1
##  9             1          0        0         0        0
## 10             1          1        0         0        0
## # i 19,990 more rows
```

Q6.(a) New Variables

For dress variable, there are two new variables introduced in the model.One is dressgreen and the other is dressred. For age variable there are also two new variables introduced, one is agesenior and the other one is ageyouth.

Q6.(b) Reference levels

The reference level for dress is black, and the reference level for age is middle.

Q7. Explanations for the logistic regression models

I think there should be 10 separate lines in this model, since we use logistic regression to each variable individually. We would have lines represent the relationships between the only continuous predictor accuracy and the categorical predictors, which are age, dress, home and jail. Age has three categories and so does dress. The last two variables each has two categories. So after all the each category has one separated line with the accuracy. So there should be 10.

Q8. Model with Interaction terms

```
# Fit a model with all second-order interaction terms
lrfit0 <- classification_lr %>%
  fit(rhbmm ~ acc+age+dress+home+jail+
        acc:age+acc:dress+acc:home+acc:jail+
        age:dress+age:home+age:jail+
        dress:home+dress:jail+
        home:jail,
      data = merry_train
      )
# Use Anova to evaluate the p-values
Anova(lrfit0$fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: rhbmm
##             LR Chisq Df Pr(>Chisq)
## acc           176.35  1  < 2.2e-16 ***
## age             1.52  2    0.46844
## dress        1354.33  2  < 2.2e-16 ***
## home         2374.87  1  < 2.2e-16 ***
## jail           15.22  1  9.573e-05 ***
## acc:age         6.20  2    0.04511 *
## acc:dress       0.95  2    0.62067
## acc:home       35.83  1  2.148e-09 ***
## acc:jail        0.32  1    0.57351
## age:dress       1.55  4    0.81697
## age:home        1.71  2    0.42497
## age:jail        0.95  2    0.62085
## dress:home    431.18  2  < 2.2e-16 ***
## dress:jail      4.21  2    0.12195
## home:jail       4.15  1    0.04158 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the output above, acc:home and dress:home meet the 99% significance level.

Q9. Backwards stepwise regression

Q9.(a) New model without Jail

```
# First fit a new model without jail interactions
lrfit1 <- classification_lr %>%
  fit(rhbmm ~ acc+age+dress+home+jail+
        acc:age+acc:dress+acc:home+
        age:dress+age:home+
        dress:home,
      data = merry_train
```

```
  )
Anova(lrfit1$fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: rhbmm
##            LR Chisq Df Pr(>Chisq)
## acc          176.83  1  < 2.2e-16 ***
## age            1.50  2    0.47237
## dress       1352.90  2  < 2.2e-16 ***
## home        2377.81  1  < 2.2e-16 ***
## jail          15.22  1  9.573e-05 ***
## acc:age        6.04  2    0.04883 *
## acc:dress      0.98  2    0.61178
## acc:home      35.46  1  2.610e-09 ***
## age:dress      1.50  4    0.82658
## age:home       1.76  2    0.41416
## dress:home   429.79  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the output above we can learn that the interaction terms acc:dress, age:dress and age:home cannot meet the 95% significance level, so we just remove them and fit a new model.

Q9.(b) Clean the interaction terms

```
# Fit a model with the interaction terms removed
lrfit2 <- classification_lr %>%
  fit(rhbmm ~ acc+age+dress+home+jail+
        acc:age+acc:home+
        dress:home,
      data = merry_train
  )
Anova(lrfit2$fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: rhbmm
##            LR Chisq Df Pr(>Chisq)
## acc          177.74  1  < 2.2e-16 ***
## age            1.48  2    0.47598
## dress       1352.46  2  < 2.2e-16 ***
## home        2376.60  1  < 2.2e-16 ***
## jail          15.26  1  9.368e-05 ***
## acc:age        5.75  2    0.05629 .
## acc:home      40.02  1  2.515e-10 ***
## dress:home   437.42  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the output above, the age variable doesn't meet 95% significance level, so we should remove it and its interactions because of the "principle of marginality" next.

Q9.(c) Clean the individual terms

```
# Fit a model with age removed
merry_lr_fit <- classification_lr %>%
  fit(rhbmm ~ acc+dress+home+jail+acc:home+dress:home,
      data = merry_train
  )
Anova(merry_lr_fit$fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: rhbmm
##             LR Chisq Df Pr(>Chisq)
## acc           178.17  1  < 2.2e-16 ***
## dress        1351.39  2  < 2.2e-16 ***
## home         2376.83  1  < 2.2e-16 ***
## jail           15.33  1  9.008e-05 ***
## acc:home       40.04  1  2.484e-10 ***
## dress:home    436.74  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now we have a model with only individual variables and interaction terms with 95% significance level.

Q10. Analysis the interactions

Q10.(a) Interaction terms

The interaction between accuracy and home and the interaction between dress and home are significant.

Q10.(b) Hypotheses

I think the archer's residence determines their accuracy of archery and their dress color. The forest archers are more likely to have a higher accuracy and dress themselves in protect colors like green. Also these interactions may affect whether they are married or not, the forest archers with higher accuracy and with green outfits dressed may have less chance to get married.

Q11. The equation for log-odds

The equation of $\hat{r}_i$ should be,

$$\hat{r}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_6 x_1 x_4 + \hat{\beta}_7 x_2 x_4 + \hat{\beta}_8 x_3 x_4$$

As the output above and for this equation,
$\hat{\beta}_0$ means the intercept with the reference under the circumstances of dressing in black, living in the city and having not been in jail before.
x_1 means the log-odds of accuracy predictors and $\hat{beta}_1$ is its slope.
x_2 means the log-odds under dressing green and $\hat{beta}_2$ is its slope.

x_3 means the log-odds of dressing in red and $\hat{beta}_3$ is its slope.
x_4 means the log-odds of living in forest and $\hat{beta}_4$ is its slope.
x_5 means the log-odds of being in jail before and $\hat{beta}_5$ is its slope.
$\hat{\beta}_6 x_1 x_4$ is the interaction term of accuracy and living in the forest with its coefficient.
$\hat{\beta}_7 x_2 x_4$ is the interaction term of dressing green and living in the forest with its coefficient.
$\hat{\beta}_8 x_3 x_4$ is the interaction term of dressing red and living in the forest with its coefficient.

Q12. Find the geometric meaning of the model

Q12.(a) Separate lines

We have 8 lines of this model. Because we have 8 terms in the equations with predictors, each term means there should be a line existing.

Q12.(b) Are they paralleled?

They are not paralleled as they have different coefficients from each other which means the slopes of every line is different.

Q13. Fit the equation with model coefficents

```
# Show the summary of the latest regression model
summary(merry_lr_fit$fit)
```

```
##
## Call:
## stats::glm(formula = rhbmm ~ acc + dress + home + jail + acc:home +
##     dress:home, family = stats::binomial, data = data)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.30221    0.10188 -12.782  < 2e-16 ***
## acc                    1.29097    0.14104   9.153  < 2e-16 ***
## dressgreen             0.89712    0.04208  21.318  < 2e-16 ***
## dressred              -0.17683    0.04213  -4.198 2.70e-05 ***
## homeforest             0.62396    0.22466   2.777  0.00548 **
## jailTRUE               0.17460    0.04464   3.911 9.19e-05 ***
## acc:homeforest         2.03914    0.32447   6.285 3.29e-10 ***
## dressgreen:homeforest  2.05219    0.23076   8.893  < 2e-16 ***
## dressred:homeforest   -1.07924    0.08603 -12.545  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27400  on 19999  degrees of freedom
## Residual deviance: 23160  on 19991  degrees of freedom
## AIC: 23178
```

```
##
## Number of Fisher Scoring iterations: 7
```

As the output above and what we described in question 11, we could put the coefficients into their positions where they should be.
So the equation could also be written like,

$$\hat{r}_i = -1.302 + 1.29x_1 + 0.897x_2 - 0.177x_3 + 0.624x_4 + 0.175x_5 + 2.04x_1x_4 + 2.05x_2x_4 - 1.079x_3x_4$$

Q14. Estimations with the model

Q14.(a) Esitimation 1

With an archer who is middle aged, lives in the forest, wears red clothing and has been to jail we should calculate with intercept plus $-0.177x_3$, $0.624x_4$, $0.175x_5$ and $-1.08x_3x_4$, with x_3, x_4, x_5 all being 1.
Which should be

$$\begin{aligned} &-1.302 - 0.177x_3 + 0.624x_4 + 0.175x_5 - 1.079x_3x_4 \\ &= -1.302 - 0.177 + 0.624 + 0.175 - 1.079 \\ &= -1.759 \end{aligned}$$

Q14.(b) Esitimation 2

With an archer who is quite old, lives in the city, usually wears black clothing and has never been to jail, we should calculate with just intercept. So it should be -1.302

Q15. Testing the model

```
# Create a new prediction tibble with true class and predicted class
merry_test_preds <- bind_cols(
  predict(merry_lr_fit, new_data = merry_test),
  truth = merry_test$rhbmm,
  predict(merry_lr_fit, new_data = merry_test, type = 'prob'),
)
merry_test_preds
```

```
## # A tibble: 10,000 x 4
##    .pred_class truth .pred_0 .pred_1
##    <fct>       <fct>   <dbl>   <dbl>
##  1 1           0       0.431   0.569
##  2 1           1       0.134   0.866
##  3 1           1       0.162   0.838
##  4 0           1       0.601   0.399
##  5 0           0       0.593   0.407
##  6 1           1       0.388   0.612
##  7 1           0       0.338   0.662
##  8 1           1       0.136   0.864
##  9 0           1       0.658   0.342
## 10 0           1       0.597   0.403
## # i 9,990 more rows
```

```r
# Show the dimensions of the tibble
dim(merry_test_preds)
```

```
## [1] 10000     4
```

Q16. Evaluating the model

Q16.(a) Confusion matrix and accuracy

Find the confusion matrix and the accuracy of the model.

```r
# Build the confusion matrix
merry_test_cm <- merry_test_preds %>%
  conf_mat(
    .pred_class,
    truth = truth
  )
merry_test_cm
```

```
##           Truth
## Prediction    0    1
##          0 3079 2058
##          1 1321 3542
```

The confusion matrix is above.

```r
# Calculate the accuracy
merry_test_preds %>% accuracy(
  .pred_class,
  truth = truth
)
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.662
```

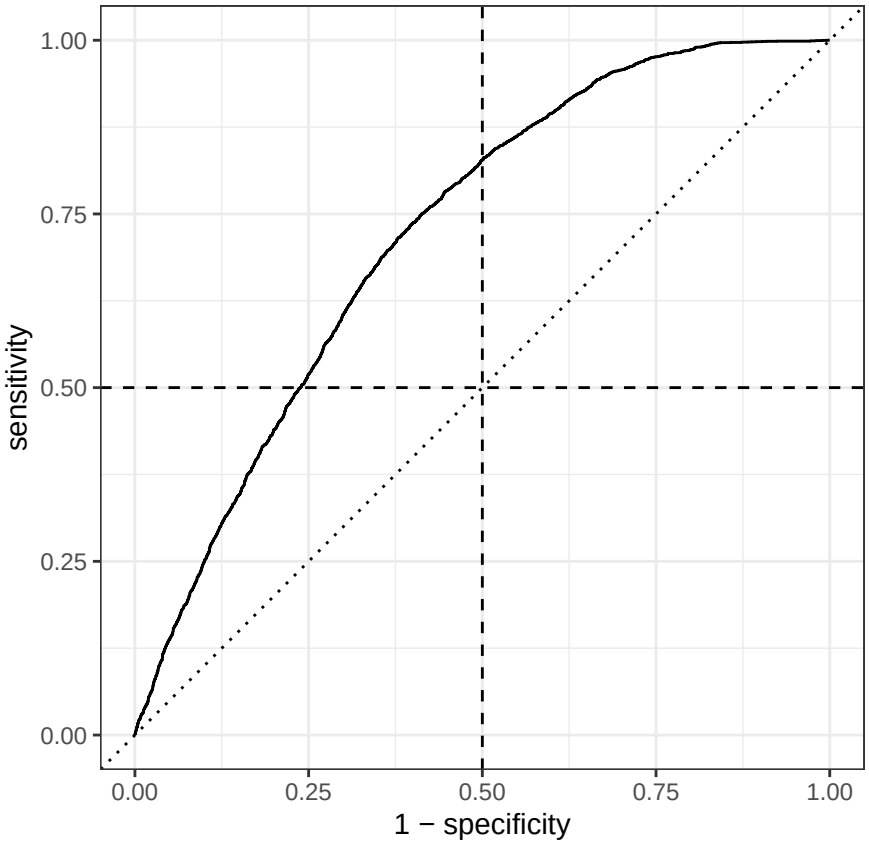
The accuracy of this model is 0.662.

Q16.(b) Sensitivity and specificity

We would calculate the sensitivity first. It should be the true positives proportion of the sum of true positives and false negatives.
Which should be $\frac{3548}{2052+3548} \approx 0.634$
Then the specificity, which should be the true negatives proportion of the sum of true negatives and false positives. Which should be $\frac{3076}{3076+1324} \approx 0.699$

Q16.(c) ROC plot

```
# Plot the roc curve with probability predictions
merry_test_preds %>%
  roc_curve(
    .pred_0,
    truth = truth
  ) %>%
  autoplot()+
  geom_vline(xintercept = 0.5, linetype = 'dashed') +
  geom_hline(yintercept = 0.5, linetype = 'dashed')
```



Q16.(d) AUC-ROC

```
merry_test_preds %>%
  roc_auc(
    .pred_0,
    truth = truth,
    event_level = 'first'
  )
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.722
```

The AUC of the ROC curve is 0.721.

Q17. Conclusions

AS we know the young lad has an accuracy of 112/116, he haven't been in jail before, he's wearing a green outfit and lives in the forest. According to the model before we could make a prediction like this,

```
# create a tibble to store the target's information
target <- tibble(
  acc=112/116,age='youth',dress='green',home='forest',jail=FALSE
)
# Predict with the model
target_preds <- bind_cols(
  target,
  predict(merry_lr_fit,new_data = target),
  predict(merry_lr_fit,new_data = target, type = 'prob')
  )
# Show the result
target_preds
```

```
## # A tibble: 1 x 8
##     acc age   dress home   jail  .pred_class .pred_0 .pred_1
##   <dbl> <chr> <chr> <chr>  <lgl> <fct>         <dbl>   <dbl>
## 1 0.966 youth green forest FALSE 1           0.00413   0.996
```

So according to the output the new archer is very likely to be a member of Robin Hood's Merry Men, with about 99.60%.