M09_activity

Hannah Valenty

2024-07-25

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
df <- read.csv('data/credit_data.csv')
head(df)
```

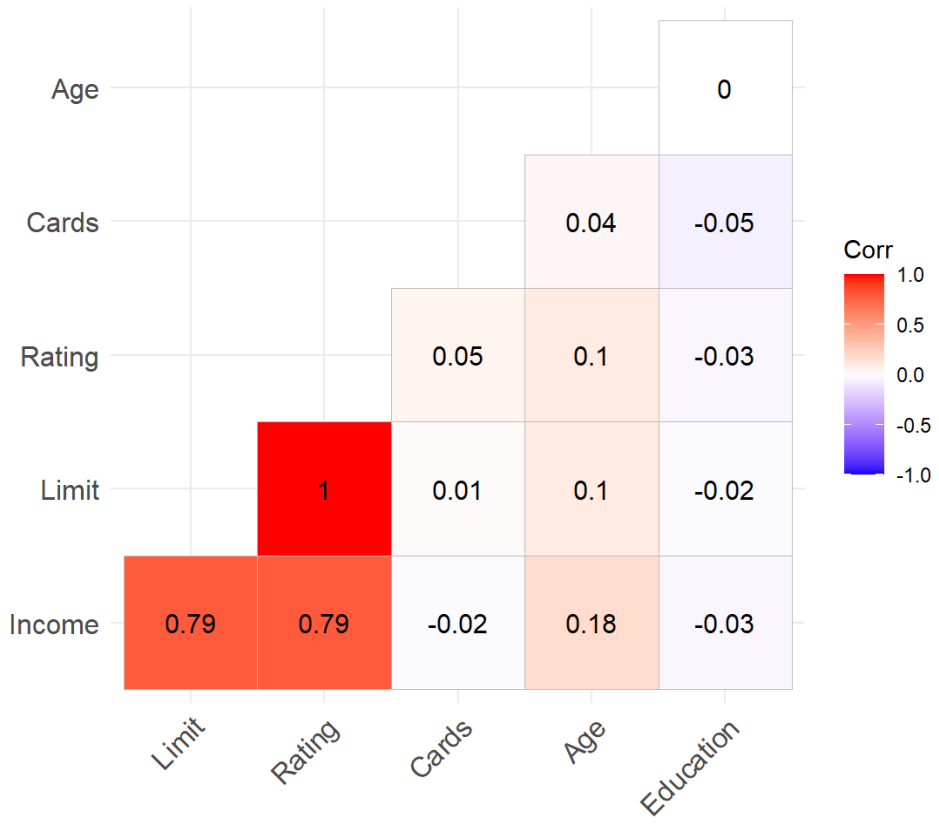
```
##   ID  Income Limit Rating Cards Age Education Gender Student Married Ethnicity
## 1  1  14.891  3606    283     2  34        11   Male      No     Yes Caucasian
## 2  2 106.025  6645    483     3  82        15 Female     Yes     Yes     Asian
## 3  3 104.593  7075    514     4  71        11   Male      No      No     Asian
## 4  4 148.924  9504    681     3  36        11 Female      No      No     Asian
## 5  5  55.882  4897    357     2  68        16   Male      No     Yes Caucasian
## 6  6  80.180  8047    569     4  77        10   Male      No      No Caucasian
##   Balance
## 1     333
## 2     903
## 3     580
## 4     964
## 5     331
## 6    1151
```

Task 1

```
library(ggcorrplot)
```

```
## Warning: package 'ggcorrplot' was built under R version 4.4.1
```

```
df2 <- df[2:7]
cor_mat <- round(cor(df2), 2)
ggcorrplot(cor_mat, lab=T, type='lower')
```



Yes, there appears to be a multicollinearity issue between limit and rating. They have a correlation of 1, which is well above the threshold of 0.8.

Task 2

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.4.1
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.4.1
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
mod0 <- lm(Balance~Income+Limit+Rating+Cards+Age+Education, data=df)
coef(mod0)
```

```
##  (Intercept)       Income        Limit       Rating        Cards          Age
## -477.9580884   -7.5580366    0.1258512    2.0631007   11.5915580   -0.8923978
##    Education
##    1.9982825
```

```
vif(mod0)
```

```
##     Income      Limit     Rating      Cards        Age  Education
##   2.773276 228.848290 230.612596   1.433932   1.038541   1.008043
```

The largest VIF is Rating (229.49), the next model will remove this.

```
mod1 <- lm(Balance~Income+Limit+Cards+Age+Education, data=df)
coef(mod1)
```

```
##  (Intercept)       Income        Limit        Cards          Age    Education
## -421.0001229   -7.4888226    0.2628007   21.6054747   -0.8907917    1.5481189
```

```
vif(mod1)
```

```
##    Income     Limit     Cards       Age Education
##  2.759808  2.699771  1.007173  1.038539  1.003564
```

This rebuilt model is more appropriate, and has all VIF values between 1 and 3 (none over 5 and especially not over 10).

Task 3.A: Prediction Interval

```
df2 <- data.frame(Income=65, Limit=6000, Cards=4,
                  Age=60, Education=10)

predict(mod1, newdata=df2, interval='prediction', level=0.95)
```

```
##        fit      lwr      upr
## 1 717.4863 396.1612 1038.811
```

With the input predictor values used for the model, the predicted Balance value has a predicted interval between 396.1612 and 1038.811, with 95% confidence.

Task 3.B: Confidence Interval

```
predict(mod1, newdata=df2, interval='confidence', level=0.95)
```

```
##        fit      lwr      upr
## 1 717.4863 689.5073 745.4653
```

We are 95% confident that the mean predicted Balance value given the input predictor values will fall within a confidence interval of 689.5073 and 745.4653. This is affirming because the interval is smaller than the predicted interval using the same parameters.