Midterm\_test.R

DavidAarhus

2020-04-14

rm(list = ls()) #removing all variables  
library("ISLR")  
library("ggplot2")  
library("corrplot")

## corrplot 0.84 loaded

library("caret")

## Loading required package: lattice

library("tidyverse")

## ── Attaching packages ─────────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ tibble 2.1.3 ✓ dplyr 0.8.3  
## ✓ tidyr 1.0.0 ✓ stringr 1.4.0  
## ✓ readr 1.3.1 ✓ forcats 0.5.0  
## ✓ purrr 0.3.3

## ── Conflicts ────────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()  
## x purrr::lift() masks caret::lift()

library("leaps")  
library("doBy")  
library("glmnet")

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

## Loaded glmnet 3.0-2

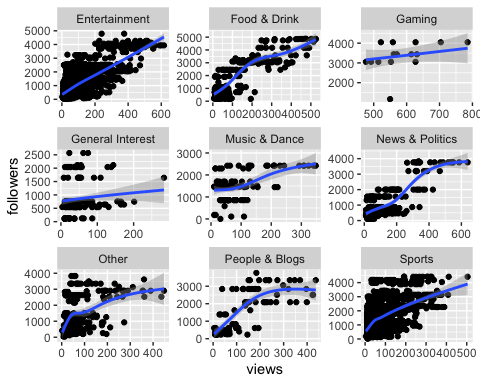
library("glmnetUtils")

##   
## Attaching package: 'glmnetUtils'

## The following objects are masked from 'package:glmnet':  
##   
## cv.glmnet, glmnet

youtube\_train <- read.csv("/Users/DavidAarhus/Documents/310 R/Datasets/youtube\_train\_A.csv")  
youtube\_test <- read.csv("/Users/DavidAarhus/Documents/310 R/Datasets/youtube\_test\_A.csv")  
  
#Question 2  
# Using ggplot, produce a scatter plot of views by the number of followers,   
# faceted by genre of the video. Include a linear trendline for each plot.   
# For which video genres is the relationship   
# between views and followers the strongest?  
  
#code  
ggplot(youtube\_train, aes(views, followers)) + geom\_point() +   
 stat\_smooth() + facet\_wrap(vars(genre), scales = "free")

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



#Question 3  
# Use summaryBy in the “doBy” package to compare average video views and duration   
# for sponsored and non-sponsored videos. Are sponsored videos lengthier than   
# non-sponsored videos? Which one has more views on average?  
  
#code  
summaryBy( views ~ sponsored, youtube\_train, FUN = mean)

## sponsored views.mean  
## 1 0 80.85023  
## 2 1 61.33790

summaryBy( duration ~ sponsored, youtube\_train, FUN = mean)

## sponsored duration.mean  
## 1 0 4.047825  
## 2 1 3.878759

#Question 4  
  
# Build a linear regression model that predicts total views as a function of   
# video duration, the number of followers, age\_18\_24, female\_percent, and type.   
# Interpret the coefficient on followers. What variables are statistically significant?  
  
model <- lm(views ~ duration   
 + followers   
 + age\_18\_24  
 + female\_percent  
 + type , youtube\_train)  
summary(model)

##   
## Call:  
## lm(formula = views ~ duration + followers + age\_18\_24 + female\_percent +   
## type, data = youtube\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -187.10 -34.21 -8.32 20.37 514.29   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8.979e+01 6.055e+00 -14.829 < 2e-16 \*\*\*  
## duration -5.296e-01 1.040e-01 -5.091 3.67e-07 \*\*\*  
## followers 6.327e-02 9.486e-04 66.694 < 2e-16 \*\*\*  
## age\_18\_24 5.126e-01 9.734e-02 5.266 1.45e-07 \*\*\*  
## female\_percent 1.374e-01 4.568e-02 3.009 0.00263 \*\*   
## typeinfluencer 8.822e+01 5.935e+00 14.863 < 2e-16 \*\*\*  
## typemedia company 8.113e+01 5.228e+00 15.517 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 68.52 on 5783 degrees of freedom  
## Multiple R-squared: 0.4609, Adjusted R-squared: 0.4604   
## F-statistic: 824.1 on 6 and 5783 DF, p-value: < 2.2e-16

#Question 5  
# Predict the number of views and calculate the RMSE (root mean squared error)   
# for the training and test sets. How effective is the model at predicting the  
# number of views? Comment on the quality of the model overall.   
# Is the model overfit or underfit, and how do we know?  
  
preds\_train1 <- predict(model, newdata = youtube\_train)  
preds\_test1 <- predict(model, newdata = youtube\_test)  
  
preds\_train1\_df <- data.frame(true = youtube\_train$views,  
 pred = predict(model, newdata = youtube\_train),  
 resid = youtube\_train$views - predict(model, newdata = youtube\_train))  
  
preds\_test1\_df <- data.frame(true = youtube\_test$views,  
 pred = predict(model, newdata = youtube\_test),  
 resid = youtube\_test$views - predict(model, newdata = youtube\_test))  
  
# MSE for train and test  
RMSE(preds\_train1\_df$pred, preds\_train1\_df$true)

## [1] 68.48125

RMSE(preds\_test1\_df$pred, preds\_test1\_df$true)

## [1] 72.40425

#Question 6  
# Estimate a forward stepwise model to predict total views using all predictors.   
# Set the maximum number of variables equal to 6. Print the stepwise model and   
# state which variables the model selects at the last step (model with 6 variables).  
  
forward\_mod <- regsubsets(views ~ .,  
 data = youtube\_train,  
 nvmax = 6,  
 method = "forward")  
summary(forward\_mod)

## Subset selection object  
## Call: regsubsets.formula(views ~ ., data = youtube\_train, nvmax = 6,   
## method = "forward")  
## 27 Variables (and intercept)  
## Forced in Forced out  
## genreFood & Drink FALSE FALSE  
## genreGaming FALSE FALSE  
## genreGeneral Interest FALSE FALSE  
## genreMusic & Dance FALSE FALSE  
## genreNews & Politics FALSE FALSE  
## genreOther FALSE FALSE  
## genrePeople & Blogs FALSE FALSE  
## genreSports FALSE FALSE  
## us\_creator FALSE FALSE  
## typeinfluencer FALSE FALSE  
## typemedia company FALSE FALSE  
## duration FALSE FALSE  
## female\_percent FALSE FALSE  
## age\_13\_17 FALSE FALSE  
## age\_18\_24 FALSE FALSE  
## age\_25\_34 FALSE FALSE  
## age\_35\_44 FALSE FALSE  
## age\_45\_54 FALSE FALSE  
## age\_55\_plus FALSE FALSE  
## us\_audience FALSE FALSE  
## sponsored FALSE FALSE  
## monthly\_views\_growth FALSE FALSE  
## followers FALSE FALSE  
## uploads\_90d FALSE FALSE  
## month FALSE FALSE  
## dow FALSE FALSE  
## live FALSE FALSE  
## 1 subsets of each size up to 6  
## Selection Algorithm: forward  
## genreFood & Drink genreGaming genreGeneral Interest genreMusic & Dance  
## 1 ( 1 ) " " " " " " " "   
## 2 ( 1 ) " " "\*" " " " "   
## 3 ( 1 ) " " "\*" " " " "   
## 4 ( 1 ) " " "\*" " " " "   
## 5 ( 1 ) " " "\*" " " " "   
## 6 ( 1 ) " " "\*" " " " "   
## genreNews & Politics genreOther genrePeople & Blogs genreSports  
## 1 ( 1 ) " " " " " " " "   
## 2 ( 1 ) " " " " " " " "   
## 3 ( 1 ) " " " " " " "\*"   
## 4 ( 1 ) " " " " " " "\*"   
## 5 ( 1 ) "\*" " " " " "\*"   
## 6 ( 1 ) "\*" " " " " "\*"   
## us\_creator typeinfluencer typemedia company duration female\_percent  
## 1 ( 1 ) " " " " " " " " " "   
## 2 ( 1 ) " " " " " " " " " "   
## 3 ( 1 ) " " " " " " " " " "   
## 4 ( 1 ) " " " " " " " " " "   
## 5 ( 1 ) " " " " " " " " " "   
## 6 ( 1 ) " " " " " " " " " "   
## age\_13\_17 age\_18\_24 age\_25\_34 age\_35\_44 age\_45\_54 age\_55\_plus  
## 1 ( 1 ) " " " " " " " " " " " "   
## 2 ( 1 ) " " " " " " " " " " " "   
## 3 ( 1 ) " " " " " " " " " " " "   
## 4 ( 1 ) "\*" " " " " " " " " " "   
## 5 ( 1 ) "\*" " " " " " " " " " "   
## 6 ( 1 ) "\*" " " " " " " " " " "   
## us\_audience sponsored monthly\_views\_growth followers uploads\_90d month  
## 1 ( 1 ) " " " " " " "\*" " " " "   
## 2 ( 1 ) " " " " " " "\*" " " " "   
## 3 ( 1 ) " " " " " " "\*" " " " "   
## 4 ( 1 ) " " " " " " "\*" " " " "   
## 5 ( 1 ) " " " " " " "\*" " " " "   
## 6 ( 1 ) " " "\*" " " "\*" " " " "   
## dow live  
## 1 ( 1 ) " " " "   
## 2 ( 1 ) " " " "   
## 3 ( 1 ) " " " "   
## 4 ( 1 ) " " " "   
## 5 ( 1 ) " " " "   
## 6 ( 1 ) " " " "

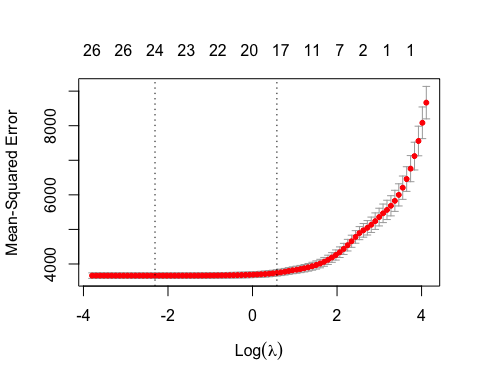
#Question 7  
# Estimate a Lasso model over the same data (using all the predictors).   
# Print the coefficients using lambda.min and lambda.1se.   
# Print the MSE plot as a function of lambda in Lasso.   
# How many variables does the model select for lambda.1se?  
  
# estimate Lasso mod   
lasso\_mod <- cv.glmnet(views ~ .,  
 data = youtube\_train,  
 alpha = 1)  
  
# print the coefficients for lambda.min and lambda.1se  
lasso\_mod$lambda.min

## [1] 0.09953851

lasso\_mod$lambda.1se

## [1] 1.780395

## Print the MSE plot as a function of lambda in Lasso.   
plot(lasso\_mod)



# put in a matrix  
coef(lasso\_mod, s = lasso\_mod$lambda.1se)

## 30 x 1 sparse Matrix of class "dgCMatrix"  
## 1  
## (Intercept) 36.37755093  
## genreEntertainment .   
## genreFood & Drink 0.52907728  
## genreGaming 301.02897291  
## genreGeneral Interest -8.27563124  
## genreMusic & Dance -10.55260503  
## genreNews & Politics 36.54965295  
## genreOther -16.60060338  
## genrePeople & Blogs .   
## genreSports -39.51458278  
## us\_creator .   
## typebrand -58.52157285  
## typeinfluencer 4.62751878  
## typemedia company .   
## duration -0.57280719  
## female\_percent -0.12261966  
## age\_13\_17 2.12308157  
## age\_18\_24 .   
## age\_25\_34 .   
## age\_35\_44 -0.31356998  
## age\_45\_54 .   
## age\_55\_plus -1.11971758  
## us\_audience -0.06135879  
## sponsored -22.41172639  
## monthly\_views\_growth 26.10794644  
## followers 0.06135040  
## uploads\_90d .   
## month .   
## dow .   
## live .

#Question 8  
# If the goal is to maximize the exposure of the videos (views),   
# what would you suggest as the best strategy to the content creators?   
# What variables affect the viewership?   
# What would you recommend as to the best model to predict views?   
# How confident would you feel about your best model?