## Model Validation

# 

library(tidyverse)

## -- Attaching packages ----------------------------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.1.0 v purrr 0.3.2   
## v tibble 2.1.1 v dplyr 0.8.0.1  
## v tidyr 0.8.3 v stringr 1.4.0   
## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts -------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

#install.packages("caret")  
library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

bike = read\_csv("hour.csv")

## Parsed with column specification:  
## cols(  
## instant = col\_double(),  
## dteday = col\_date(format = ""),  
## season = col\_double(),  
## yr = col\_double(),  
## mnth = col\_double(),  
## hr = col\_double(),  
## holiday = col\_double(),  
## weekday = col\_double(),  
## workingday = col\_double(),  
## weathersit = col\_double(),  
## temp = col\_double(),  
## atemp = col\_double(),  
## hum = col\_double(),  
## windspeed = col\_double(),  
## casual = col\_double(),  
## registered = col\_double(),  
## count = col\_double()  
## )

bike = bike %>% mutate(season = as\_factor(as.character(season))) %>%  
 mutate(season = fct\_recode(season,  
 "Spring" = "1",  
 "Summer" = "2",  
 "Fall" = "3",  
 "Winter" = "4"))  
bike = bike %>% mutate(yr = as\_factor(as.character(yr))) %>%  
 mutate(mnth = as\_factor(as.character(mnth))) %>%  
 mutate(hr = as\_factor(as.character(hr)))  
bike = bike %>% mutate(holiday = as\_factor(as.character(holiday))) %>%  
 mutate(holiday = fct\_recode(holiday,  
 "NotHoliday" = "0",  
 "Holiday" = "1"))  
  
bike = bike %>% mutate(workingday = as\_factor(as.character(workingday))) %>%  
 mutate(workingday = fct\_recode(workingday,  
 "NotWorkingDay" = "0",  
 "WorkingDay" = "1"))  
  
bike = bike %>% mutate(weathersit = as\_factor(as.character(weathersit))) %>%  
 mutate(weathersit = fct\_recode(weathersit,  
 "NoPrecip" = "1",  
 "Misty" = "2",  
 "LightPrecip" = "3",  
 "HeavyPrecip" = "4"))  
  
bike = bike %>% mutate(weekday = as\_factor(as.character(weekday))) %>%  
 mutate(weekday = fct\_recode(weekday,  
 "Sunday" = "0",  
 "Monday" = "1",  
 "Tuesday" = "2",  
 "Wednesday" = "3",  
 "Thursday" = "4",  
 "Friday" = "5",  
 "Saturday" = "6"))

set.seed(1234)  
train.rows = createDataPartition(y=bike$count, p=0.7, list=FALSE)  
train = bike[train.rows,]  
test = bike[-train.rows,]  
  
nrow(train)

## [1] 12167

nrow(test)

## [1] 5212

There are 12,167 rows in the train data set and 5,212 in the test data set.

mod1=lm(count ~ season + mnth + hr + holiday + weekday + temp + weathersit, train )  
summary(mod1)

##   
## Call:  
## lm(formula = count ~ season + mnth + hr + holiday + weekday +   
## temp + weathersit, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -411.57 -62.29 -9.66 51.54 494.52   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -87.1390 6.9960 -12.456 < 2e-16 \*\*\*  
## seasonSummer 34.0014 6.3399 5.363 8.33e-08 \*\*\*  
## seasonFall 27.1663 7.4964 3.624 0.000291 \*\*\*  
## seasonWinter 60.2453 6.3962 9.419 < 2e-16 \*\*\*  
## mnth2 0.6289 5.1046 0.123 0.901951   
## mnth3 7.4480 5.7452 1.296 0.194867   
## mnth4 -6.6612 8.5213 -0.782 0.434401   
## mnth5 -6.2329 9.1424 -0.682 0.495407   
## mnth6 -15.8184 9.3673 -1.689 0.091306 .   
## mnth7 -39.2578 10.4561 -3.755 0.000174 \*\*\*  
## mnth8 -21.7608 10.2226 -2.129 0.033300 \*   
## mnth9 1.3338 9.0877 0.147 0.883319   
## mnth10 0.9570 8.4836 0.113 0.910185   
## mnth11 -15.1008 8.1639 -1.850 0.064382 .   
## mnth12 -12.2448 6.4726 -1.892 0.058542 .   
## hr1 -13.3293 6.9652 -1.914 0.055682 .   
## hr2 -27.4480 7.0006 -3.921 8.87e-05 \*\*\*  
## hr3 -33.8591 7.0797 -4.783 1.75e-06 \*\*\*  
## hr4 -37.7544 7.1298 -5.295 1.21e-07 \*\*\*  
## hr5 -20.8072 7.0678 -2.944 0.003247 \*\*   
## hr6 37.4750 7.0673 5.303 1.16e-07 \*\*\*  
## hr7 174.5062 6.9408 25.142 < 2e-16 \*\*\*  
## hr8 310.6002 7.0497 44.059 < 2e-16 \*\*\*  
## hr9 172.3560 7.0135 24.575 < 2e-16 \*\*\*  
## hr10 112.8882 7.0375 16.041 < 2e-16 \*\*\*  
## hr11 139.8538 7.0762 19.764 < 2e-16 \*\*\*  
## hr12 182.1016 7.0797 25.722 < 2e-16 \*\*\*  
## hr13 177.8863 7.0168 25.351 < 2e-16 \*\*\*  
## hr14 163.2828 7.1329 22.891 < 2e-16 \*\*\*  
## hr15 178.1201 7.0976 25.096 < 2e-16 \*\*\*  
## hr16 231.1350 7.1679 32.246 < 2e-16 \*\*\*  
## hr17 382.4767 7.0346 54.371 < 2e-16 \*\*\*  
## hr18 361.1422 7.1736 50.343 < 2e-16 \*\*\*  
## hr19 237.1363 7.0249 33.757 < 2e-16 \*\*\*  
## hr20 166.4963 6.9865 23.831 < 2e-16 \*\*\*  
## hr21 114.6982 6.9704 16.455 < 2e-16 \*\*\*  
## hr22 75.1763 7.0002 10.739 < 2e-16 \*\*\*  
## hr23 35.4147 6.9890 5.067 4.10e-07 \*\*\*  
## holidayHoliday -21.8882 6.4894 -3.373 0.000746 \*\*\*  
## weekdaySunday -16.5691 3.7640 -4.402 1.08e-05 \*\*\*  
## weekdayMonday -7.9035 3.8915 -2.031 0.042277 \*   
## weekdayTuesday -7.1190 3.7953 -1.876 0.060717 .   
## weekdayWednesday -7.4042 3.7927 -1.952 0.050938 .   
## weekdayThursday -0.9102 3.7787 -0.241 0.809662   
## weekdayFriday -0.3409 3.7732 -0.090 0.928011   
## temp 288.5138 12.1631 23.721 < 2e-16 \*\*\*  
## weathersitMisty -19.1163 2.3603 -8.099 6.06e-16 \*\*\*  
## weathersitLightPrecip -90.5259 3.7350 -24.237 < 2e-16 \*\*\*  
## weathersitHeavyPrecip 83.0764 111.2351 0.747 0.455166   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 111 on 12118 degrees of freedom  
## Multiple R-squared: 0.6229, Adjusted R-squared: 0.6214   
## F-statistic: 417.1 on 48 and 12118 DF, p-value: < 2.2e-16

Using the variables season, mnth, hr, holiday, weekday, temp and weathersit the datasets R squared value is relatively high which means the model quality is fairly good. Its difficult to rely on the p-values in this dataset due to how large it is and how many variables are being used but the temp and most of the hr variables are the more significant variables in this model.

predict\_train = predict(mod1,newdata = train)  
head(predict\_train, 6)

## 1 2 3 4 5 6   
## -36.99526 -51.11404 -51.75482 -55.65016 -57.81925 13.80902

The first 6 predictions for count of bike rides are mostly negative which suggests that the model is underfitting the data. It would be unrealistic for there to be a negative count of bike rides.

mod2=lm(count ~ season + mnth + hr + holiday + weekday + temp + weathersit, test )  
summary(mod2)

##   
## Call:  
## lm(formula = count ~ season + mnth + hr + holiday + weekday +   
## temp + weathersit, data = test)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -391.56 -62.05 -8.62 53.04 500.99   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -81.6755 10.4473 -7.818 6.47e-15 \*\*\*  
## seasonSummer 38.5716 9.7193 3.969 7.33e-05 \*\*\*  
## seasonFall 24.9370 11.5940 2.151 0.031534 \*   
## seasonWinter 77.5316 9.6737 8.015 1.35e-15 \*\*\*  
## mnth2 3.3479 7.9172 0.423 0.672414   
## mnth3 0.6466 8.8525 0.073 0.941773   
## mnth4 -2.7938 13.1935 -0.212 0.832305   
## mnth5 -3.1799 13.9724 -0.228 0.819981   
## mnth6 -21.6182 14.4251 -1.499 0.134026   
## mnth7 -45.0398 16.4572 -2.737 0.006226 \*\*   
## mnth8 -17.5068 15.7902 -1.109 0.267608   
## mnth9 12.6603 14.0432 0.902 0.367351   
## mnth10 -15.2650 12.8942 -1.184 0.236523   
## mnth11 -25.9159 12.4594 -2.080 0.037572 \*   
## mnth12 -22.9579 9.8833 -2.323 0.020223 \*   
## hr1 -30.3465 10.8283 -2.803 0.005090 \*\*   
## hr2 -24.8723 10.7989 -2.303 0.021305 \*   
## hr3 -46.5344 10.7437 -4.331 1.51e-05 \*\*\*  
## hr4 -48.7347 10.6064 -4.595 4.43e-06 \*\*\*  
## hr5 -33.3872 10.5692 -3.159 0.001593 \*\*   
## hr6 24.6648 10.4841 2.353 0.018680 \*   
## hr7 157.1585 10.9373 14.369 < 2e-16 \*\*\*  
## hr8 310.8457 10.4668 29.698 < 2e-16 \*\*\*  
## hr9 146.4236 10.5959 13.819 < 2e-16 \*\*\*  
## hr10 109.2998 10.5591 10.351 < 2e-16 \*\*\*  
## hr11 138.1318 10.5351 13.112 < 2e-16 \*\*\*  
## hr12 175.4796 10.6307 16.507 < 2e-16 \*\*\*  
## hr13 172.5028 11.0037 15.677 < 2e-16 \*\*\*  
## hr14 154.9559 10.6638 14.531 < 2e-16 \*\*\*  
## hr15 150.3653 10.8434 13.867 < 2e-16 \*\*\*  
## hr16 231.1447 10.5595 21.890 < 2e-16 \*\*\*  
## hr17 389.9368 10.9411 35.640 < 2e-16 \*\*\*  
## hr18 334.0384 10.3558 32.256 < 2e-16 \*\*\*  
## hr19 252.3557 10.7367 23.504 < 2e-16 \*\*\*  
## hr20 149.0260 10.7782 13.827 < 2e-16 \*\*\*  
## hr21 100.5719 10.7818 9.328 < 2e-16 \*\*\*  
## hr22 66.8151 10.6486 6.275 3.79e-10 \*\*\*  
## hr23 29.0448 10.6646 2.723 0.006482 \*\*   
## holidayHoliday -35.0499 9.4091 -3.725 0.000197 \*\*\*  
## weekdaySunday -13.7894 5.7630 -2.393 0.016759 \*   
## weekdayMonday -7.1115 5.9235 -1.201 0.229977   
## weekdayTuesday -5.6614 5.8058 -0.975 0.329535   
## weekdayWednesday 4.0747 5.7703 0.706 0.480126   
## weekdayThursday -5.3475 5.8104 -0.920 0.357440   
## weekdayFriday 6.1834 5.7659 1.072 0.283579   
## temp 285.6045 18.9416 15.078 < 2e-16 \*\*\*  
## weathersitMisty -20.1322 3.6555 -5.507 3.82e-08 \*\*\*  
## weathersitLightPrecip -90.9184 6.0098 -15.128 < 2e-16 \*\*\*  
## weathersitHeavyPrecip -159.3746 79.5565 -2.003 0.045199 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 111.9 on 5163 degrees of freedom  
## Multiple R-squared: 0.6313, Adjusted R-squared: 0.6279   
## F-statistic: 184.2 on 48 and 5163 DF, p-value: < 2.2e-16

predict\_test = predict(mod2,newdata = test)  
head(predict\_test, 6)

## 1 2 3 4 5 6   
## -13.13044 156.14153 159.27390 213.75796 202.07321 15.78093

The first 6 predictions using the test data set seem more realistic since the predictions are positive values.

SSE = sum((test$count - predict\_test)^2)  
SST = sum((test$count - mean(test$count))^2)  
1 - SSE/SST

## [1] 0.6312857

The manually calculated R squared value is very comparable to the training data set. The manually calculated R squared value was 0.63 and the R squared value in the training model was 0.62. Both values indicate that the model quality is good.

The major difference between k fold cross validation and model validation via a training/testing split is that the k fold method leaves enough data for training and validation. The train/test split method reduces the data used in the training portion which leaves greater risk for losing important trends which increases potential error. When the k fold method is used, the data is divided into different subsets, each time being a different subset that is left out and then the rror is averaged over all the different trials. By doing the k fold method, the risk of error is significantly reduced because the majority of the data that is being used in the subsetting is also used in the validating process.