Problem Set #1

Bennett Hellman
15.072 - Advanced Analytics Edge
MIT

September 22, 2021

1. (a) i. In-sample: RMSE = 1582.19, MAE = 1259.63, $R^2 = 0.83$
Out-of-sample: RMSE = 6300.77, MAE = 4694.05, $OSR^2 = 0.47$

   (b) i. Based on the p-values in the previous model, I eliminated CPI.All because of its statistical insignificance. This is essentially one iteration of backwards selection. I terminated the process, however, I could consider eliminating CPI.Energy because of its high correlation with Wrangler.Queries.

   ii. In-sample: RMSE = 1597.04, MAE = 1256.77, $R^2 = 0.82$
Out-of-sample: RMSE = 6016.04, MAE = 4270.67, $OSR^2 = 0.52$

   iii. As expected, the sparser model performed worse in sample. However, The sparser model provided in problem b provided slightly better out-of-sample RMSE, MAE, and OSR^2. This is likely because it is less fit to the training data.

   (c) Note: I treated month as a categorical variable.

   i. In-sample: RMSE 1103.22, MAE = 865.97, $R^2 = 0.92$
Out-of-sample: RMSE = 1512.56, MAE = 1283.01, $OSR^2 = 0.88$

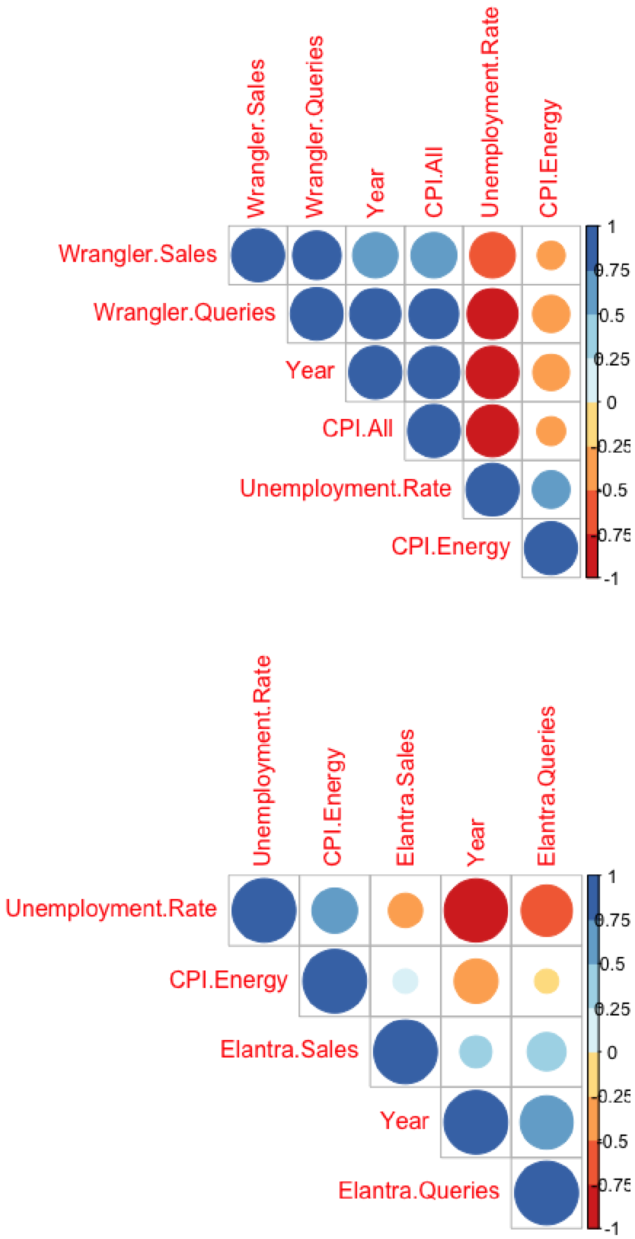
   ii. The model accounting for seasonality provides definitively better in and out-of-sample performance metrics compared to the model in problem a. This is likely because sales fluctuate with season since because everybody wants to be driving door-less Wrangler in the summertime.

   (d) i. In sample: RMSE = 2795.83, MAE = 2191.22, $R^2 = 0.64$
Out of sample: RMSE = 4102.87, MAE = 3230.09, OSR^2 = -2.21

   (e) i. Although the model performs well for wrangler sales the analogous model performs poorly for Elantra sales.

   ii. Figure 1 demonstrates the variables in the Wrangler model have higher correlation with Wrangler Sales than Elantra model variables do with Elantra Sales. This is demonstrated by the smaller and less bold colors in the second graphic. In layman's terms, these variables individually do not show patterns with Elantra.Sales, so we would likely not expect them collectively to explain Elantra.Sales well.

Figure 1: Wrangler Model Variance vs. Elantra Model Variance



References

I used code from a few open source websites, as cited in my code.

```
#Admin
setwd("/Users/bennetthellman/Desktop/OneDrive - Massachusetts Institute of
Technology/AE")
df <- read.csv("WranglerElantra2018.csv", stringsAsFactors = FALSE, header
= TRUE)
library(tidyverse)
library(miscTools)
library(Metrics)


#Data Wrangling
df<-df%>%mutate(Month = as.factor(Month))

#a.i
#set.seed(15072)
#row.number <- sample(1:nrow(df), 0.7*nrow(df))
#train = df[row.number,]
#test = df[-row.number,]

train = df%>%filter(Year != 2018)
test = df%>%filter(Year == 2018)


#a.i
lmai_train =
lm(Wrangler.Sales~Year+Unemployment.Rate+Wrangler.Queries+CPI.Energy+CPI.All,
data = train) #Create the linear regression
summary(lmai_train)
predai_train <- predict(lmai_train, newdata = train)
c(RMSE = rmse(train$Wrangler.Sales, predai_train), MAE =
mae(train$Wrangler.Sales, predai_train), R2=rSquared(train$Wrangler.Sales,
resid = train$Wrangler.Sales-predai_train))
predai <- predict(lmai_train, newdata = test)
train.mean <- mean(train$Wrangler.Sales)
SSE <- sum((predai - test$Wrangler.Sales)^2)
SST <- sum((train.mean - test$Wrangler.Sales)^2)
OSR2 <- 1 - SSE/SST
c(RMSE = rmse(test$Wrangler.Sales, predai), MAE = mae(test$Wrangler.Sales,
predai), OSR2 = OSR2)


#b.i
cov(df[,c('Unemployment.Rate', 'Wrangler.Queries', 'CPI.Energy',
'CPI.All')])
summary(lmai_train)
#b.ii
lmbii_train =
lm(Wrangler.Sales~Year+Unemployment.Rate+Wrangler.Queries+CPI.Energy, data
= train) #Create the linear regression
summary(lmbii_train)
predbi_train <- predict(lmbii_train, newdata = train)
```

```
c(RMSE = rmse(train$Wrangler.Sales, predbi_train), MAE =
mae(train$Wrangler.Sales, predbi_train), R2=rSquared(train$Wrangler.Sales,
resid = train$Wrangler.Sales-predbi_train))
predbii <- predict(lmbii_train, newdata = test)
train.mean <- mean(train$Wrangler.Sales)
SSE <- sum((predbii - test$Wrangler.Sales)^2)
SST <- sum((train.mean - test$Wrangler.Sales)^2)
OSR2 <- 1 - SSE/SST
c(RMSE = rmse(test$Wrangler.Sales, predbii), MAE = mae(test$Wrangler.Sales,
predbii), OSR2 = OSR2)
#b.iii


#c.i
lmci_train =
lm(Wrangler.Sales~Year+Unemployment.Rate+Wrangler.Queries+CPI.Energy+CPI.All+Month,
data = train) #Create the linear regression
summary(lmci_train)
predci_train <- predict(lmci_train, newdata = train)
c(RMSE = rmse(train$Wrangler.Sales, predci_train), MAE =
mae(train$Wrangler.Sales, predci_train), R2=rSquared(train$Wrangler.Sales,
resid = train$Wrangler.Sales-predci_train))
predci <- predict(lmci_train, newdata = test)
train.mean <- mean(train$Wrangler.Sales)
SSE <- sum((predci - test$Wrangler.Sales)^2)
SST <- sum((train.mean - test$Wrangler.Sales)^2)
OSR2 <- 1 - SSE/SST
c(RMSE = rmse(test$Wrangler.Sales, predci), MAE = mae(test$Wrangler.Sales,
predci), OSR2 = OSR2)
#c.ii

#d.i
lmdi_train = lm(Elantra.Sales~ Year + Unemployment.Rate + Elantra.Queries +
CPI.Energy + CPI.All + Month, data = train) #Create the linear regression
summary(lmdi_train)
preddi_train <- predict(lmdi_train, newdata = train)
c(RMSE = rmse(train$Elantra.Sales, preddi_train), MAE =
mae(train$Elantra.Sales, preddi_train), R2=rSquared(train$Elantra.Sales,
resid = train$Elantra.Sales-preddi_train))
preddi <- predict(lmdi_train, newdata = test)
train.mean <- mean(train$Elantra.Sales)
SSE <- sum((preddi - test$Elantra.Sales)^2)
SST <- sum((train.mean - test$Elantra.Sales)^2)
OSR2 <- 1 - SSE/SST
c(RMSE = rmse(test$Elantra.Sales, preddi), MAE = mae(test$Elantra.Sales,
preddi), OSR2 = OSR2)

#e.i

#e.ii

library(corrplot)
library(RColorBrewer)
```

```
df_wrang<-
df%>%select(Wrangler.Sales,Year,Unemployment.Rate,Wrangler.Queries,CPI.Energy,CPI.All)
M <-cor(df_wrang)
corrplot(M, type="upper", order="hclust",
         col=brewer.pal(n=8, name="RdYlBu"))

df_elantra<-df%>%select(Elantra.Sales,
Year,Unemployment.Rate,Elantra.Queries,CPI.Energy)
M <-cor(df_elantra)
corrplot(M, type="upper", order="hclust",
         col=brewer.pal(n=8, name="RdYlBu"))


#http://www.sthda.com/english/wiki/correlation-analyses-in-r
```