Robert Final StockX Sale Prediction

Robert Carter

5/2/2021

library(readxl)  
library(ISLR)  
library(tree)  
stockx\_df <- read\_excel('/Users/robertcarter/Desktop/StockX-Data-Contest-2019-3.xlsx')  
shoedata <- as.data.frame(stockx\_df)  
  
head(shoedata)

## Order Date Brand Sneaker Name Sale Price  
## 1 2017-09-01 Yeezy Adidas-Yeezy-Boost-350-Low-V2-Beluga 1097  
## 2 2017-09-01 Yeezy Adidas-Yeezy-Boost-350-V2-Core-Black-Copper 685  
## 3 2017-09-01 Yeezy Adidas-Yeezy-Boost-350-V2-Core-Black-Green 690  
## 4 2017-09-01 Yeezy Adidas-Yeezy-Boost-350-V2-Core-Black-Red 1075  
## 5 2017-09-01 Yeezy Adidas-Yeezy-Boost-350-V2-Core-Black-Red-2017 828  
## 6 2017-09-01 Yeezy Adidas-Yeezy-Boost-350-V2-Core-Black-Red-2017 798  
## Retail Price Release Date Shoe Size Buyer Region  
## 1 220 2016-09-24 11.0 California  
## 2 220 2016-11-23 11.0 California  
## 3 220 2016-11-23 11.0 California  
## 4 220 2016-11-23 11.5 Kentucky  
## 5 220 2017-02-11 11.0 Rhode Island  
## 6 220 2017-02-11 8.5 Michigan

Renaming the fields to make it easier when calling the fields back.

names(shoedata) <- tolower(gsub("\\.", "\_", names(shoedata)))  
names(shoedata) <- gsub("\\(", "", names(shoedata))  
names(shoedata) <- gsub("\\)", "", names(shoedata))  
names(shoedata) <- gsub(" ", "\_", names(shoedata))

Taking a look at the data. We are looking at the names that were just changed. Also, took a look at the unique brands of the data and there were only 2. Last, did a quick summary of the data. It look like the highest retail price for the shoes at 250, the highest sale price was 4050, and the median sale price was 370. It looks like there is a good chance that by puchasing a pair of shoes will lead to making a profit.

names(shoedata)

## [1] "order\_date" "brand" "sneaker\_name" "sale\_price" "retail\_price"  
## [6] "release\_date" "shoe\_size" "buyer\_region"

unique(shoedata$brand)

## [1] "Yeezy" "Off-White"

summary(shoedata)

## order\_date brand sneaker\_name   
## Min. :2017-09-01 00:00:00 Length:99956 Length:99956   
## 1st Qu.:2018-05-02 00:00:00 Class :character Class :character   
## Median :2018-09-24 00:00:00 Mode :character Mode :character   
## Mean :2018-08-12 21:24:50   
## 3rd Qu.:2018-12-15 00:00:00   
## Max. :2019-02-13 00:00:00   
## sale\_price retail\_price release\_date   
## Min. : 186.0 Min. :130.0 Min. :2015-06-27 00:00:00   
## 1st Qu.: 275.0 1st Qu.:220.0 1st Qu.:2017-09-09 00:00:00   
## Median : 370.0 Median :220.0 Median :2017-12-16 00:00:00   
## Mean : 446.6 Mean :208.6 Mean :2018-02-10 04:24:17   
## 3rd Qu.: 540.0 3rd Qu.:220.0 3rd Qu.:2018-08-03 00:00:00   
## Max. :4050.0 Max. :250.0 Max. :2019-02-07 00:00:00   
## shoe\_size buyer\_region   
## Min. : 3.500 Length:99956   
## 1st Qu.: 8.000 Class :character   
## Median : 9.500 Mode :character   
## Mean : 9.344   
## 3rd Qu.:11.000   
## Max. :17.000

Doing a simple linear regression on the data. Checking to see how significant brand is to the data. Brand appears to be statistically signifacnt to the data set, the Yeey brand in particular.

slr <- lm(sale\_price ~ brand, data = shoedata)  
summary(slr)

##   
## Call:  
## lm(formula = sale\_price ~ brand, data = shoedata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -468.5 -100.0 -49.0 48.0 3378.5   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 671.481 1.287 521.6 <2e-16 \*\*\*  
## brandYeezy -311.449 1.515 -205.6 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 214.6 on 99954 degrees of freedom  
## Multiple R-squared: 0.2972, Adjusted R-squared: 0.2972   
## F-statistic: 4.226e+04 on 1 and 99954 DF, p-value: < 2.2e-16

We are opening up the linear regression model to more field. This time we are checking the significance of brand, release date, retail price, and shoe size compared to sale price. Again, all fields are significant.

lmshoes <- lm(sale\_price ~ brand + release\_date + retail\_price + shoe\_size, data=shoedata)  
summary(lmshoes)

##   
## Call:  
## lm(formula = sale\_price ~ brand + release\_date + retail\_price +   
## shoe\_size, data = shoedata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -469.8 -111.8 -4.9 46.7 3293.3   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.249e+03 5.316e+01 117.550 < 2e-16 \*\*\*  
## brandYeezy -3.784e+02 2.097e+00 -180.417 < 2e-16 \*\*\*  
## release\_date -3.689e-06 3.385e-08 -108.978 < 2e-16 \*\*\*  
## retail\_price 2.407e-01 3.679e-02 6.542 6.12e-11 \*\*\*  
## shoe\_size 2.322e+00 2.764e-01 8.402 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 202.3 on 99951 degrees of freedom  
## Multiple R-squared: 0.3752, Adjusted R-squared: 0.3752   
## F-statistic: 1.501e+04 on 4 and 99951 DF, p-value: < 2.2e-16

Going for further into data exploratoin, now taking a look athe the means for both brands. The first shoes the mean sale price for both brands. The Off-White brands on average resale s higher than Yeezy. However, on average Yeezy retail for more than Off-White.

tapply(shoedata$sale\_price, shoedata$brand, mean)

## Off-White Yeezy   
## 671.4812 360.0326

tapply(shoedata$retail\_price, shoedata$brand, mean)

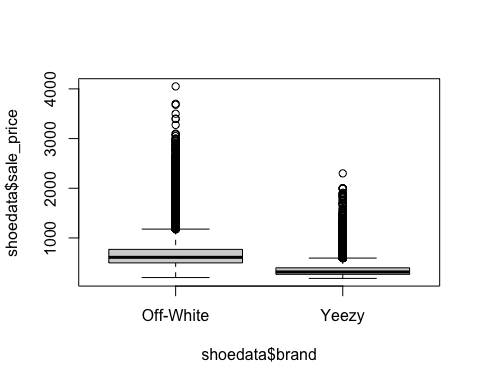
## Off-White Yeezy   
## 179.3783 219.8739

Now we are going to take a look at box plots for both mean of sale price and retail price for each brand.

boxplot(shoedata$retail\_price ~ shoedata$brand)



boxplot(shoedata$sale\_price ~ shoedata$brand)

 Does brand affect resale price H0:The mean will not be different between brnads HA:The mean will be diferent between brands For tis test we are going to take a look a one way anova test to see if there is any significane in the mean for sale price between brands. The is high significance in the means, therefore we do not reject the null hypothesis.

oneway.test(shoedata$sale\_price~shoedata$brand)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: shoedata$sale\_price and shoedata$brand  
## F = 22476, num df = 1, denom df = 31821, p-value < 2.2e-16

Double check the one way test results with an anova test and Tukey multiple pairwise-comparisons. The results are again significant and the Tukey comparison shows a $311 difference between sale price.

aov.out = aov(sale\_price~brand, data=shoedata)  
summary(aov.out)

## Df Sum Sq Mean Sq F value Pr(>F)   
## brand 1 1.946e+09 1.946e+09 42261 <2e-16 \*\*\*  
## Residuals 99954 4.603e+09 4.606e+04   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(aov.out)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = sale\_price ~ brand, data = shoedata)  
##   
## $brand  
## diff lwr upr p adj  
## Yeezy-Off-White -311.4486 -314.418 -308.4792 0

Creating new data set, the original data set. Going to predict try to predict sale price that is higher than the mean sale price.

shoedata2 <- shoedata

Creating a variable that determines if the sale price is higher than the average sale price. Doing this to determine the categories that best predict sale price

shoedata2$high\_sale <- ifelse(shoedata2$sale\_price > 371, 'Yes', 'No')  
shoedata2$high\_sale <- as.factor(shoedata2$high\_sale)

Creating a decision tree based on high sale

tree.shoedata2 = tree(high\_sale~.-sale\_price, data=shoedata2)

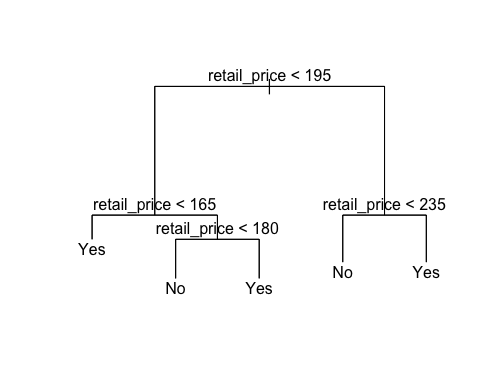
## Warning in tree(high\_sale ~ . - sale\_price, data = shoedata2): NAs introduced by  
## coercion

summary(tree.shoedata2)

##   
## Classification tree:  
## tree(formula = high\_sale ~ . - sale\_price, data = shoedata2)  
## Variables actually used in tree construction:  
## [1] "retail\_price"  
## Number of terminal nodes: 5   
## Residual mean deviance: 1.038 = 103700 / 99950   
## Misclassification error rate: 0.2885 = 28841 / 99956

Plotting the decision tree. Looking at the tree shoes with a retail price under 165 has a good chance of having a high resale price. If retail price is greater than 180 then the resle price will be high and if retail price is greater than 235 the resale price will be high.

plot(tree.shoedata2)  
text(tree.shoedata2, pretty = 0)



Breaking the data into train and test data set. Also, going to shuffle the data. It is probably ok without shuffling but it doesn’t hurt.

#This is where the shuffle happens. This generates a random list of the data. Mixing the data will be important in our decision tree.  
set.seed(101)  
shuffle\_index <- sample(1:nrow(shoedata2))  
head(shuffle\_index)

## [1] 2873 19665 87391 35772 46326 61056

#Creating the new dataset with the mixed up data set.  
shoedata2<- shoedata2[shuffle\_index, ]  
  
#Breaking the data into train and test datasets. Splitting the data 80/20  
smp\_siz = floor(0.8\*nrow(shoedata2))  
train\_ind = sample(seq\_len(nrow(shoedata2)),size = smp\_siz)   
train =shoedata2[train\_ind,] #creates the training dataset with row numbers stored in train\_ind  
test=shoedata2[-train\_ind,]

Predicting high sale price. First train the model on the train data set. Next will run the model using test data set. We were able to acurately predict high sale price about 42% of the time. However, we were able to predict no high sale price about 99% of the time.

tree.train = tree(high\_sale~.-sale\_price, data=train)

## Warning in tree(high\_sale ~ . - sale\_price, data = train): NAs introduced by  
## coercion

tree.pred = predict(tree.train, test, type="class")

## Warning in pred1.tree(object, tree.matrix(newdata)): NAs introduced by coercion

with(test, table(tree.pred, high\_sale))

## high\_sale  
## tree.pred No Yes  
## No 10098 5744  
## Yes 2 4148