Avocados – Relation between   
Type, Prices and Volume

Sai Sri Sindhuja Pragadaraju

J02227522

Johnson and Wales University,

Providence. RI

Table of Contents

[Introduction 3](#_Toc37679758)

[Data collecting and Attributes 3](#_Toc37679759)

[Data Exploration 4](#_Toc37679760)

[Data cleaning and wrangling 4](#_Toc37679761)

[Patterns and Feature selection 5](#_Toc37679762)

[Logistic Regression 6](#_Toc37679763)

[Naïve Bayes 8](#_Toc37679764)

[Decision tree 10](#_Toc37679765)

[Neural Network 12](#_Toc37679766)

[Conclusion and future scope 14](#_Toc37679767)

[References 16](#_Toc37679768)

[Appendix 17](#_Toc37679769)

[Code 17](#_Toc37679770)

# Introduction

This project is taken as an extension of one of the academic projects done in the past to deep drive into the information to explore new results and pattern analysis, which can be used to predict better and improved results. The dataset is taken from HASS Avocado Board (HAB) which has the data for their HASS Avocados data for USA and other countries. Avocadoes have been associated to healthy lifestyle since the beginning of 21st century when the Latin dishes found their way into the mainstream food industry around the world, mainly into America. This fruit plays a significant role within the low fat trends which consists of Keto, Paleo, Vegan, Dukan, Zone and many more different kinds of diet routines. The prices have spiked tremendously and now the World Avocado Industry is at a whopping 11.8 Billion USD and expected to touch 15 Billion USD in the next 5 years.

The consumption in USA alone is 2.6 Billion USD in 2016 alone (Avocados, 2018). According to global reports USA imports 52% of avocados from countries like Mexico, Peru, Chile and other countries with Mexico being the major exporter with 85% of the total import. California produces 86% of the US produced avocados and HASS variety holds 95% of the California production. (Zahniser, 2018)

# Data collecting and Attributes

The dataset is collected from Kaggle website (Kiggins, 2019) which was originally taken from HASS Avocado Board official website as multiple datasets and combined. It contains of actual sales data from January 2015 to March 2018. The data is of Hass avocados sold to various regions in the United States.

There are 18249 observation and 14 attributes which are explained below.

X : UID

Date : Date of transaction

AveragePrice: Average price of avocado on that date

Total.Volume: Total Volume of avocados sold

X4046 : small sized avocados sold

X4225 : medium sized avocados sold

X4770 : large sized avocados sold

Total.Bags : Total number of bags sold

Small.Bags : small sized bags sold

Large.Bags : large sized bags sold

XLarge.Bags : Extra-large sized bags sold

type : Conventional, organic-type of produce

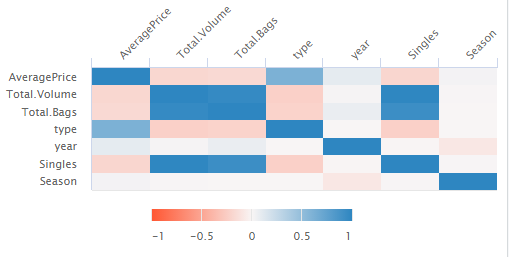
year : year of the sale

region : Region or city the sale done

# Data Exploration

## Data cleaning and wrangling

There are no missing values in the provided dataset. Date column was first separated into day and month. Then the months were grouped into seasons as follows. Season – winter (1), spring (2), summer (3), fall (4). X4046, X4225 and X4770 sized avocados sold single and hence the total values grouped into a new column Singles. Singles + Total.Bags = Total.Volume. In the project we are trying to understand the relation between the type of avocado, avg. price of the avocado, its influence by Seasons and the trend in the years. Other attributes have been removed for easy of handling, also the correlation of those attributes being low.



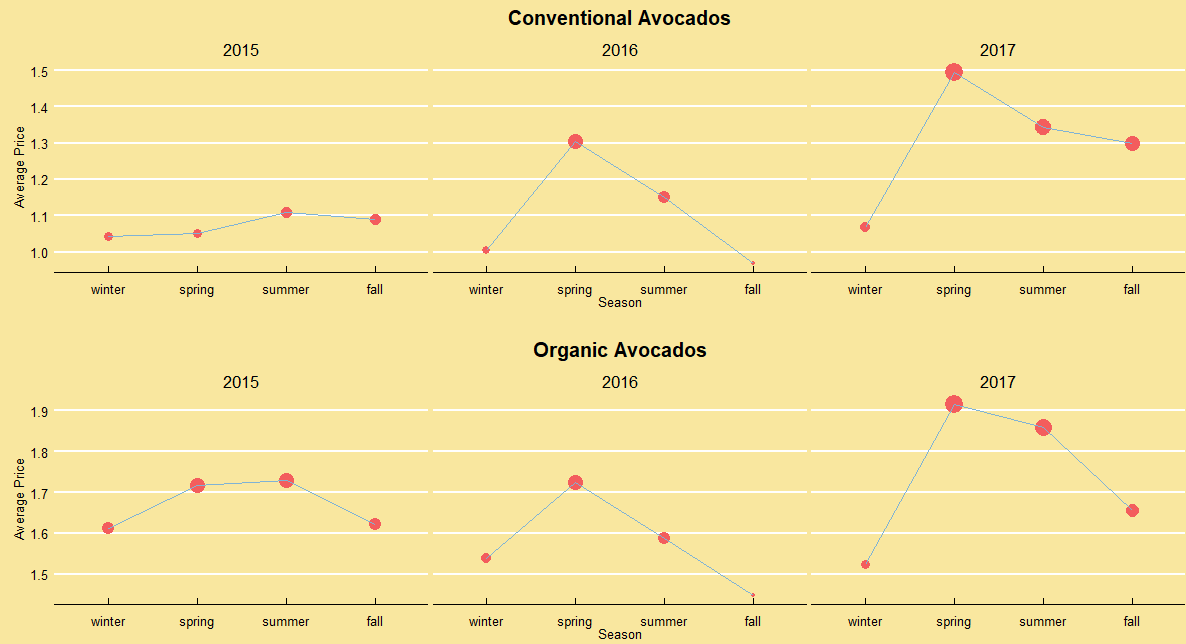
For the attribute type- Average prices is shown to have positive effect whereas the Total Volume, Total bags and single have a negative effect. Season and year do not seem to have any effects. But the total bags and singles combined gives us total volume hence these two attributes can be discarded for the selection.

**Ytype = b0 + b1\*AveragePrice – b2\*Total.Volume + error**

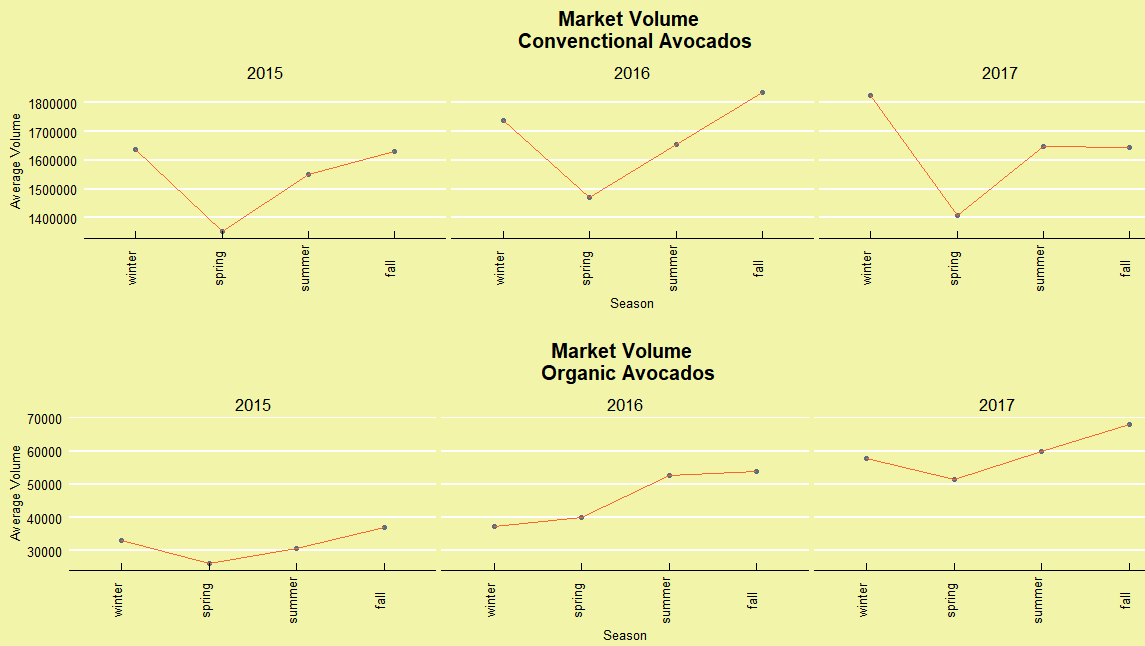
This could be the regression formula for the calculation/prediction of fruit type.

## Patterns and Feature selection

The initial analysis shows the average price trend for conventional and organic avocados. Also the effect of Season for changes in average prices and total volumes can be seen below. The average prices during Spring season have seen a jump in 2016 and 2017 for both Conventional and Organic types and drop in value from summer to fall season. This could be due to the riping period of the fruit.



To further understand, below is the total volume consumption during each season. The total volume consumed is mostly inversely proportional to the average prices of the fruit especially for Conventional type. The drop in production can be one of the many reasons for this.

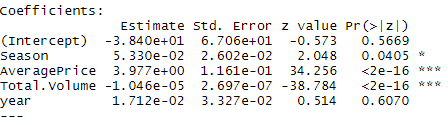


This shows that fruit type may be dependent on Average price, Total volume, Season and maybe on year. Hence, let us choose these five attributes for the model building and analysis.

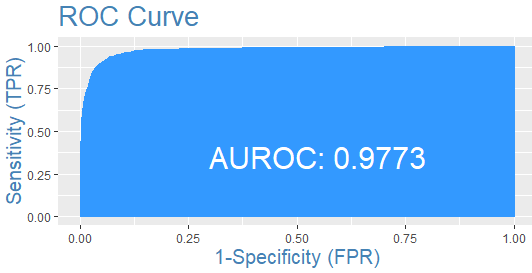
**Model Creation**

# Logistic Regression

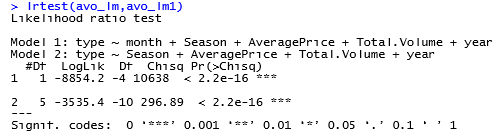
As the variable type is a binomial variable-conventional (0) or organic (1), logistic regression can be performed and its ROC, Log Likelihood, Logloss and AUC can be calculated and compared to find the best model for predictions. The results are as shown.

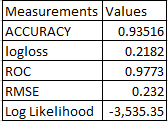


The Total.Volume attribute and Average Price have most significant impact on the fruit type whereas Season and year have less to negligible effect. The area under ROC close above 95% which is excellent. It means the type of fruit is well predicted based on the given attributes.



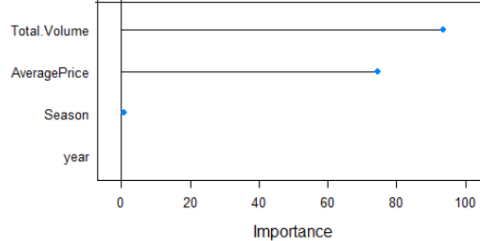
Different models produced more or less similar results, but when month variable was added there was a huge jump in the log likelihood even though the p value was <0.0001 for both. This was tested using Chi-square goodness of fit test and the final model was chosen for further comparison.



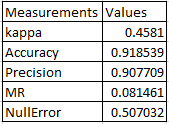
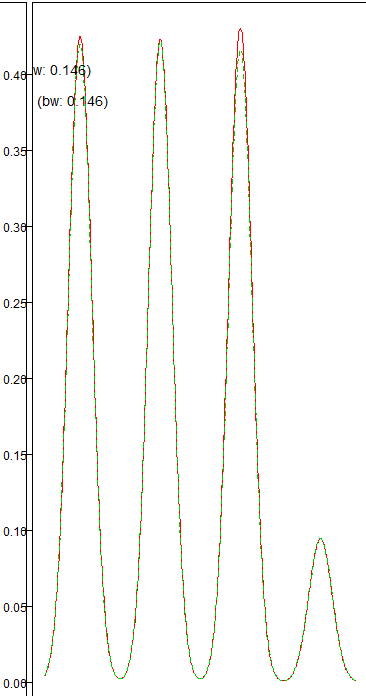


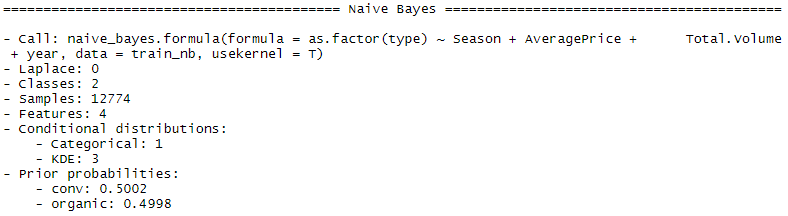
# Naïve Bayes

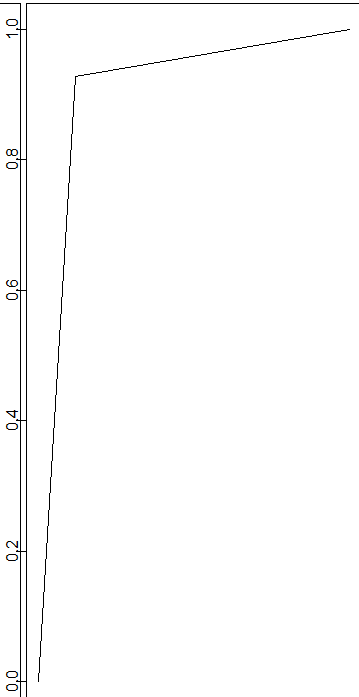
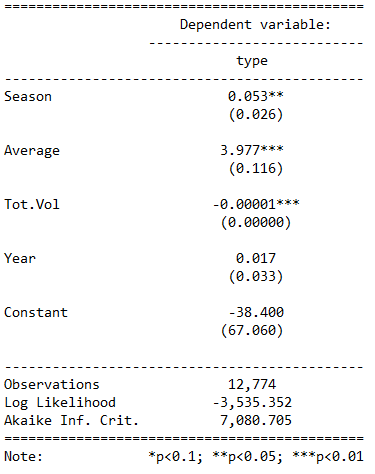
Type is taken in a binary form where 0 is conventional and 1 is organic. Naïve Bayes package using kernel based density as it gives better and accurate results compared to the traditional method. The importance of each attribute is measured and Total Volume and Average price have the most importance while season and year have least or zero.



Training and test sets both have shown accuracy above 90% and the prediction model has an accuracy of 91.8%. The error rates have also been low, making it a good model. The priori probability is equal, i.e., 50% for conventional and organic type.



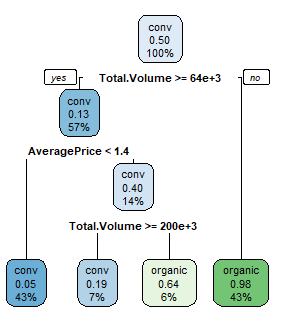




Similar to the logistic regression we have total volume and average prices with the most significant effect on the fruit type. Other algorithms are also further discussed.

# Decision tree

Decision trees are best for both classification and regression. A tree structure helps us better understand the logic behind the algorithm calculations and finding the importance of variables. The same four input variables such as Season, Average price, Total Volume and year are given to the model with single output variable as type. Below is the decision tree formed with 3depths and 4 outcomes/leaves.



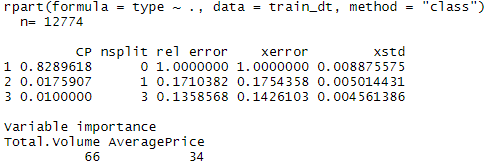
Step1. The first level is the overall probability of the type being conventional. 50% is conventional.

Step2. The Total volume if greater than 63810.68 or not is checked. If yes then 57% of the times it is total volume with 13% probability of conventional type.

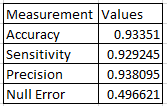
If not there is a 43% of being less than 63k volume with 98% change for being organic.

Step3. The 3rd node checks if the average price is less than 1.4. If yes, there is a 43% of it having average price less than 1.4 of which 5% changes of being conventional.

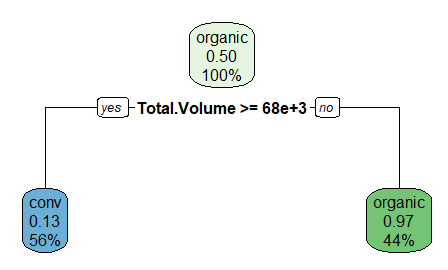
Step4. If average price greater than 1.4 then again the total volume is checked for greater than or equal to 200475.3. If yes, then chances of being organic is 6% and conventional is 7%.



As shown below the variable importance for total volume is 66% and average price is 34 percent, where season and year have none. The residuals were normally distributed hence a good fit. The results are matching with the previous model classifications.



Pruning the tree may have reduced the nodes and leaves but has not contributed much to the overall model improvement, but has reduced the Accuracy of the results.



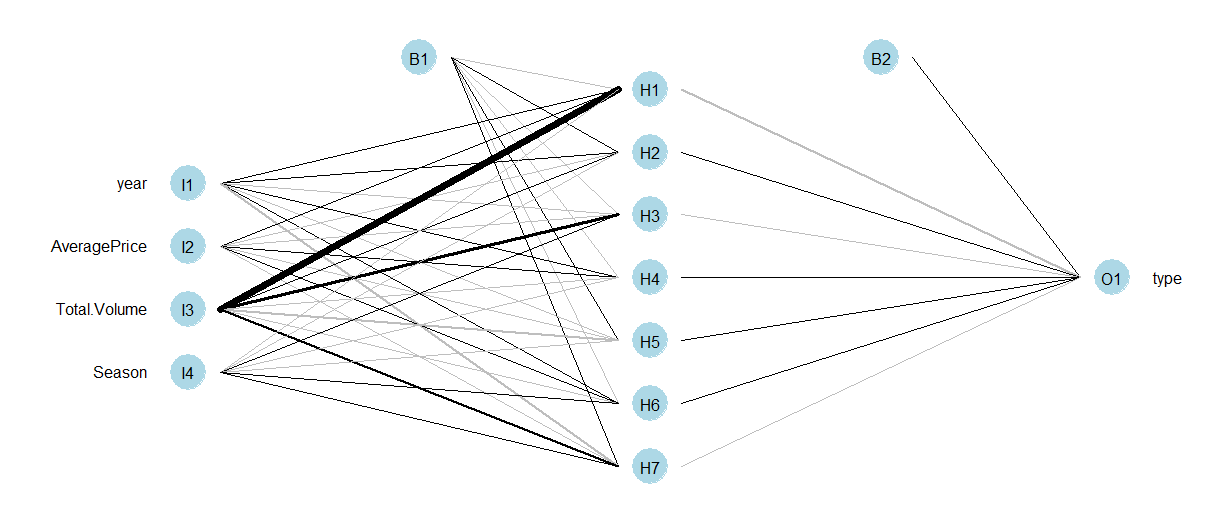
Pruning the tree should help in reducing the nodes and increasing the decision making accuracy, but that does not happen every time.

# Neural Network

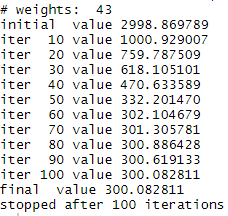
For running neural networks on the data provided, the main steps are to scale and normalize the data. That is to bring the mean close to 0. Here we have used the sigmoid function for normalization so the values fall between 0 and 1.

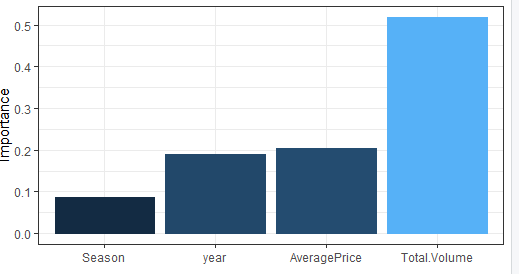
The general expression for neural networks is

**Y = ∑ (weights\*inputs) + bias**

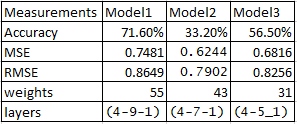


From the above figure we can understand that the weights for Total Volume (neuron) is more over other neurons of input layer. The variable importance can be seen below with total Volume having the highest importance followed by Average price, year and Season. Here the model is a 4-7-1 network meaning 4 inputs, 7 hidden layers and 1 output. The bias B1 and B2 help is the activation of the neurons by adjusting the output with the sum of weights. Total weights is 43 which meaning it need 42 inputs.

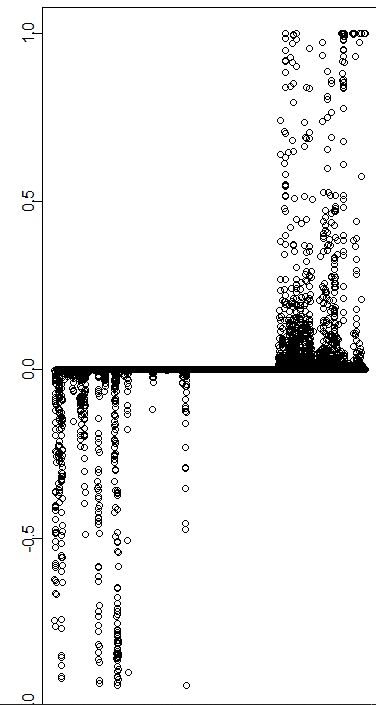




The weights decide the importance of a variable in the model. Also the RSME, MSE, Accuracy, loss function and cost function help in deciding the model validity.



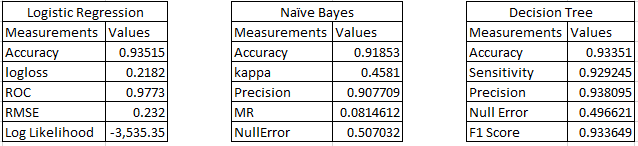
The models with hidden layers 9 and 5 had results completely different from the previous algorithm model results. The model with (4-7-1) layers could be considered as final model but has a large RMSE value which is not considered a good fitted model even though the residuals are normally distributed.



# Conclusion and future scope

It can be safely concluded that Average prices and Total Volume are strong attributes enough to predict the type of fruit. Of all the four algorithms used the models using Logistic Regression, Naïve Bayes and Decision trees had an accuracy rate of above 90%, whereas Neural Network had a very low accuracy of 33%.

Hence, the algorithms Logistic Regression, Naïve Bayes and Decision tree can be considered best for predicting the fruit type. Further deep learning can be done to improve the predictions where the dates and months of the transactions can be considered for time series forecasting to validate the hypothesis that only Total Volume and Average prices decide the fruit type prediction. The below tables shows the measurements and values for the three algorithm based model results.



Linear regression can also be done for this kind of dataset, where not just the type but also the average prices for the next day or month can be measured.

Regional analysis can also help understand the differences in average prices during various seasons. Western coast may have lesser prices compared to Eastern regions as California is the base production region for these Hass avocados.

**Limitations**

There are various limitations to this as there was shortage of Mexican avocado production in 2016 which could have resulted in the spiked rates or volume consumptions (Williams, 2016).

Also for prediction of the year 2018, the California Heat wave which has burnt the crops needs to be taken into consideration (Avocado Workers Strike Over Export Scheme, 2018).

As only Hass avocados are considered for this study, we may not be able to fully understand the US avocado consumption and average prices of the fruit while we exclude the Florida produce and imports from Mexico and other major contributing countries.

References

*Avocado Workers Strike Over Export Scheme*. (2018, November 12). Retrieved from foodlogistics.com: https://www.foodlogistics.com/risk-compliance/news/21031834/avocado-workers-strike-over-export-scheme

*Avocados*. (2018, October). Retrieved from agmrc.org: https://www.agmrc.org/commodities-products/fruits/avocados

*Decision Trees in R using rpart*. (2014, August 24). Retrieved from gormanalysis.com: https://www.gormanalysis.com/blog/decision-trees-in-r-using-rpart/

Hou, J. (2017, August 30). *Simple Neural Networks Exercises*. Retrieved from RPubs-by RStudio: https://rpubs.com/JanpuHou/303002

Khan, R. (2017, December 20). *Naive Bayes Classifier: theory and R example*. Retrieved from RPubs-by RStudio: https://rpubs.com/riazakhan94/naive\_bayes\_classifier\_e1071

Kiggins, J. (2019, 09). *Avocado Prices*. Retrieved from kaggle.com: https://www.kaggle.com/neuromusic/avocado-prices/download

Last Name, F. M. (Year). *Book Title.* City Name: Publisher Name.

tutorials, P. a. (2018, October 9). *neuralnet: Train and Test Neural Networks Using R*. Retrieved from https://datascienceplus.com/: https://datascienceplus.com/neuralnet-train-and-test-neural-networks-using-r/

Williams, G. (2016, Oct 31). *How The Avocado Shortage Is Affecting Chipotle, Grocers And You*. Retrieved from forbes.com: https://www.forbes.com/sites/geoffwilliams/2016/10/31/how-the-avocado-shortage-is-affecting-chipotle-grocers-and-you/#558ab9c55475

Zahniser, S. (2018, November 19). *U.S. and Mexican avocado production is concentrated in a small number of States*. Retrieved from United States Department of Agriculture-Economic Research Service: https://www.ers.usda.gov/data-products/chart-gallery/gallery/chart-detail/?chartId=90622

# Appendix

## Code

#packages needed

library(data.table)

library(Hmisc)

library(pastecs)

library(psych)

library(nnet)

library(foreign)

library(ggplot2)

library(reshape2)

library(FactoMineR)

library(factoextra)

library(randomForest)

library(caret)

library(quantmod)

library(neuralnet)

library(MASS)

library(VGAM)

library(kernlab)

library(mlogit)

require(caTools)

library(e1071)

require(leaps)

library(missForest)

library(mice)

library(Hmisc)

require(ggplot2)

require(GGally)

require(reshape2)

require(lme4)

require(compiler)

require(parallel)

require(boot)

require(lattice)

library(Information)

library(survival)

library(survminer)

library(dplyr)

library(readr)

library(foreign)

library(sqldf)

library(readxl)

library(Rfast)

library(lubridate)

library(lintr)

library(zoo)

library(odbc)

library(DBI)

library(RODBC)

library(xlsx)

library(funModeling)

library(tidyverse)

library(Hmisc)

library(readxl)

library(skimr)

library(fastDummies)

library(readr)

library(cowplot)

library(Amelia)

library(tidyr)

library(neuralnet)

library(NeuralNetTools)

library(nnet)

#Loading data into R

avo <- read.csv("avocado-prices/avocado11.csv")

str(avo) #structure of the dataframe

dim(avo) #dimensions of the dataframe

##Check for null

sapply(avo, function(x) sum(is.na(x)))

avo<-separate(avo,Date,into = c("date","month"))

avo2<-avo[,-c(1,2)]

avo$Season <- factor(avo2$Season, levels = c(1,2,3,4),

labels = c("winter","spring","summer","fall"))

missmap(avo2)

avo2$Season<-as.numeric(avo2$Season)

avo2\_s$month<-as.factor(avo2\_s$month)

hchart(cor(avo2[,-c(1,6:8)]))

########EDA################

basic\_eda <- function(avo)

{glimpse(avo)

df\_status(avo)

freq(avo)

profiling\_num(avo)

plot\_num(avo)

describe(avo)}

basic\_eda(avo)

#seasonal graph

ggplot(avo, aes(x = AveragePrice, fill = as.factor(year))) +

geom\_density(alpha = .5)+facet\_wrap(~ year) + theme(plot.title=element\_text(hjust=0.5),plot.background=element\_rect(fill="#F9E79F")) +

guides(fill = FALSE) + labs(title="Distribution of Prices by year", x = 'Average Price', y = 'Density')

# Detecting seasonality patterns

conv\_patterns <- avo %>% select(Season, year,AveragePrice, type) %>%

filter(type == "0",year==c("2015","2016","2017")) %>%group\_by(year,Season) %>%

summarize(avg=mean(AveragePrice)) %>%

ggplot(aes(x=Season, y=avg)) +geom\_point(color="#F35D5D", aes(size=avg)) +

geom\_line(group=1, color="#7FB3D5")+

facet\_wrap(~as.factor(year))+ theme\_economist()+

theme(legend.position="none", plot.title=element\_text(hjust=0.5),plot.background=element\_rect(fill="#F9E79F")) +

labs(title="Conventional Avocados", x="Season", y="Average Price")

org\_patterns <- avo2%>% select(year,AveragePrice, type) %>%

filter(type == "1",year==c("2015","2016","2017")) %>%

group\_by(year) %>% summarize(avg=mean(AveragePrice)) %>%

ggplot(aes(x=year, y=avg)) + geom\_point(color="#F35D5D", aes(size=avg)) +

geom\_line(group=1, color="#58D68D")+

facet\_wrap(~as.factor(year))+ theme\_economist()+

theme(legend.position="none", plot.title=element\_text(hjust=0.5),plot.background=element\_rect(fill="#F9E79F")) +

labs(title="Organic Avocados", x="year", y="Average Price")

plot\_grid(conv\_patterns, org\_patterns, nrow=2) #plotting

###########Logistic Regression #############

str(avo)

avo2$month<-as.factor(avo2$month)

## 70%:30% ratio of the sample size

smp\_size <- floor(0.70 \* nrow(avo\_nn))

## set the seed to make your partition reproducible

library(MLmetrics)

train\_ind <- sample(seq\_len(nrow(avo\_nn)), size = smp\_size)

train <- avo\_nn[train\_ind, ]

test <- avo\_nn[-train\_ind, ]

avo\_lm <- glm(type~.,

family = binomial(link = "logit"), data = train)

avo\_lm1 <- glm(type~ Season +AveragePrice+Total.Volume+year,

family = binomial(link = "logit"), data = train)

summary(avo\_lm)

confint(avo\_lm)

## Model Calculations

exp(cbind(OR = coef(avo\_lm), confint(avo\_lm)))

predicted <- plogis(predict(avo\_lm, test))

p\_lr <- predict(avo\_lm, test, type = "response")

p\_lr <- ifelse(p\_lr > 0.5,1,0)

prr<-prediction(p\_lr,test$type)

perf <- performance(prr,measure = "tpr",x.measure = "fpr")

plot(perf)

auc(test$type,p\_lr)

library(InformationValue)

optCutOff <- optimalCutoff(test$type, predicted)[1]

misClassError(test$type, predicted, threshold = optCutOff)

plotROC(test$type, predicted)

Concordance(test$type, predicted)

roc(test$type, predicted)

LogLoss(y\_pred = avo\_lm$fitted.values, y\_true = train$type)

AUC(y\_pred = avo\_lm$fitted.values, y\_true = train$type)

sensitivity(test$type, predicted, threshold = optCutOff)

specificity(test$type, predicted, threshold = optCutOff)

l\_cm <- confusionMatrix(predicted,test$type, threshold = optCutOff)

l\_cm <- table(predicted,test$type)

l\_cm

Accuracy(predicted,testset$type)

accuracy\_Test <- sum(diag(l\_cm)) / sum(l\_cm)

RMSE(test$type,predicted)

#results into text

library(stargazer)

library(apaTables)

library(MBESS)

stargazer(avo\_lm, type="text",

dep.var.labels=c("type"),

covariate.labels=c("month","Season","Average","Tot.Vol","Year"),

out="C:/Users/Sindhu/Desktop/OptStim/opt-final/models.txt")

stargazer(avo\_lm, type="text",

dep.var.labels=c("type"),

covariate.labels=c("Season","Average","Tot.Vol","Year"),

out="C:/Users/Sindhu/Desktop/OptStim/opt-final/models1.txt")

predict(avo\_lm, test, type="response")[1]

anova(avo\_lm,avo\_lm1,test = "Chisq")

library(lmtest)

lrtest(avo\_lm,avo\_lm1)

##########Decision Tree Model################

library(rpart)

library(rpart.plot)

avo\_dt<-avo2

avo\_dt$type <- factor(avo\_dt$type, levels = c(0,1), labels = c("conv", "organic"))

avo\_dt<-avo\_dt[,-c(1,5:8)]

## 70%:30% ratio of the sample size

smp\_size\_dt <- floor(0.70 \* nrow(avo\_dt))

str(avo\_dt)

## set the seed to make your partition reproducible

train\_in\_dt <- sample(seq\_len(nrow(avo\_dt)), size = smp\_size\_dt)

train\_dt <- avo\_dt[train\_in\_dt, ]

test\_dt <- avo\_dt[-train\_in\_dt, ]

dt <- rpart(type ~.,

data = train\_dt, method = 'class')

summary(dt)

rpart.plot(dt,extra=106)

p\_dt <- predict(dt, test\_dt, type = 'class')

table\_mat <- table(test\_dt$type, p\_dt)

table\_mat

accuracy\_Test <- sum(diag(table\_mat)) / sum(table\_mat)

printcp(dt)

dt$variable.importance

#prune the tree

dt1 <- prune(dt, cp = .01)

rpart.plot(dt1,extra=10)

p\_dt1 <- predict(dt1, test\_dt, type = 'class')

table\_mat <- table(test\_dt$type, p\_dt1)

table\_mat

accuracy\_Test <- sum(diag(table\_mat)) / sum(table\_mat)

printcp(dt1)

#########Naive Bayes###################

library(mice)

library(randomForest)

library(klaR)

#Setting outcome variables as categorical

avo\_nb<-avo

avo\_nb$type <- factor(avo\_nb$type, levels = c(0,1), labels = c("conv", "organic"))

str(avo\_nb)

missmap(avo\_nb)

avo\_nb<-avo\_nb[,-c(1:3,8:10)]

#creating training and test set

smp\_size <- floor(0.70 \* nrow(avo\_nb))

train\_ind <- sample(seq\_len(nrow(avo\_nb)), size = smp\_size)

train\_nb <- avo\_nb[train\_ind, ]

test\_nb <- avo\_nb[-train\_ind, ]

prop.table(table(avo\_nb$type)) \* 100

prop.table(table(train\_nb$type)) \* 100

prop.table(table(test\_nb$type)) \* 100

library(naivebayes)

#Model creation using the package naivebayes

NB\_model2<- naive\_bayes(as.factor(type) ~ Season+AveragePrice+Total.Volume + year,

usekernel = T, data=train\_nb)

NB\_model2

printALL=function(NB\_model2){

trainPred=predict(NB\_model2, newdata = train\_nb, type = "class")

trainTable=table(train\_nb$type, trainPred)

testPred=predict(NB\_model2, newdata=test\_nb, type="class")

testTable=table(test\_nb$type, testPred)

trainAcc=(trainTable[1,1]+trainTable[2,2])/sum(trainTable)

testAcc=(testTable[1,1]+testTable[2,2])/sum(testTable)

message("Contingency Table for Training Data")

print(trainTable)

message("Contingency Table for Test Data")

print(testTable)

message("Accuracy")

print(round(cbind(trainAccuracy=trainAcc, testAccuracy=testAcc),3))

}

printALL(NB\_model2)

summary(NB\_model2)

plot(NB\_model2)

p\_nb <- predict(NB\_model2, test\_nb)

p\_nb <- table(test\_nb$type, p\_nb)

p\_nb

xxx<-varImp(model\_nb)

plot(xxx)

x = train\_nb[,-c(4,5)]

y = train\_nb$type

model\_nb = train(x,y,'nb',trControl=trainControl(method='cv',number=10))

model\_nb

##############Neural Network#####################

#Loading data into R

avo\_nn <- read.csv("avocado-prices/avocado11.csv")

str(avo\_nn) #structure of the dataframe

dim(avo\_nn) #dimensions of the dataframe

##Data cleaning and wrangling

sapply(avo\_nn, function(x) sum(is.na(x)))

avo\_nn<-separate(avo\_nn,Date,into = c("date","month"))

names(avo\_nn)

avo\_nn<-avo\_nn[,-c(1,2,7:10)]

#Data Scaling

scalingavo<-avo\_nn

#scalingavo<-cbind(scalingavo,avo$Season)

str(scalingavo)

#names(scalingavo)[names(scalingavo) == "avo$Season"] <- "Season"

scalingavo$year<-as.numeric(scalingavo$year)

scalingavo$date<-as.numeric(scalingavo$date)

scalingavo$month<-as.numeric(scalingavo$month)

scalingavo<-scale(scalingavo)

#Normalisation

max = apply(scalingavo , 2 , max)

min = apply(scalingavo, 2 , min)

scaled\_avo <-as.data.frame(scale(scalingavo, center = min, scale = max - min))

#Training and Test Data

trainset <- scaled\_avo[1:12774, ]

testset <- scaled\_avo[12775:18249, ]

#Neural Network Model

nn1 = nnet(type ~ year+AveragePrice+Total.Volume+Season,

data=trainset, size=7)

print(nn1)

plotnet(nn1)

garson(nn1)

neuralweights(nn1)

#import function from Github

library(devtools)

library(MLmetrics)

source\_url('https://gist.github.com/fawda123/7471137/raw/

cd6e6a0b0bdb4e065c597e52165e5ac887f5fe95/nnet\_plot\_update.r')

wts <- neuralweights(nn1)

struct <- wts$struct

wts <- unlist(wts$wts)

plotnet(wts, struct = struct)

#Model Predictions and Checks

nn1.predict<-predict(nn1,testset)

str(nn1.predict)

RMSE(testset$type,nn1.predict)

MSE(nn1.predict,testset$type)

cm = table(testset$type, nn1.predict)

accuracy\_Test <- sum(diag(cm)) / sum(cm)

plot(nn1$residuals)

plot(nn1$wts)

qqnorm(nn1$residuals)

#########END##################

