Project

#Read in the data.   
candy<-read.csv("Candy - Data.csv", header = TRUE)  
head(candy)

## competitorname chocolate fruity caramel peanutyalmondy nougat  
## 1 100 Grand 1 0 1 0 0  
## 2 3 Musketeers 1 0 0 0 1  
## 3 One dime 0 0 0 0 0  
## 4 One quarter 0 0 0 0 0  
## 5 Air Heads 0 1 0 0 0  
## 6 Almond Joy 1 0 0 1 0  
## crispedricewafer hard bar pluribus sugarpercent pricepercent winpercent  
## 1 1 0 1 0 0.732 0.860 66.97173  
## 2 0 0 1 0 0.604 0.511 67.60294  
## 3 0 0 0 0 0.011 0.116 32.26109  
## 4 0 0 0 0 0.011 0.511 46.11650  
## 5 0 0 0 0 0.906 0.511 52.34146  
## 6 0 0 1 0 0.465 0.767 50.34755

#Creating Linear Regression model.  
lnmodel<-lm(candy$winpercent ~ . -competitorname, data = candy)  
#This gives the coefficient values.  
summary(lnmodel)

##   
## Call:  
## lm(formula = candy$winpercent ~ . - competitorname, data = candy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.2244 -6.6247 0.1986 6.8420 23.8680   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 34.5340 4.3199 7.994 1.44e-11 \*\*\*  
## chocolate 19.7481 3.8987 5.065 2.96e-06 \*\*\*  
## fruity 9.4223 3.7630 2.504 0.01452 \*   
## caramel 2.2245 3.6574 0.608 0.54493   
## peanutyalmondy 10.0707 3.6158 2.785 0.00681 \*\*   
## nougat 0.8043 5.7164 0.141 0.88849   
## crispedricewafer 8.9190 5.2679 1.693 0.09470 .   
## hard -6.1653 3.4551 -1.784 0.07852 .   
## bar 0.4415 5.0611 0.087 0.93072   
## pluribus -0.8545 3.0401 -0.281 0.77945   
## sugarpercent 9.0868 4.6595 1.950 0.05500 .   
## pricepercent -5.9284 5.5132 -1.075 0.28578   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.7 on 73 degrees of freedom  
## Multiple R-squared: 0.5402, Adjusted R-squared: 0.4709   
## F-statistic: 7.797 on 11 and 73 DF, p-value: 9.504e-09

This residual standard error is the square root of the variance. Standard error is very similar.

This R-squared subtracts the residual error from the variance.

Adjusted R-Squared normalizes Multiple R-Squared by considering how many samples you have and how many variables you are using

#Creating Logistic Regression model.  
lgmodel<-glm(candy$chocolate~.-competitorname,family = "binomial", data = candy)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

#This give the summary of the coeffienti values.  
summary(lgmodel)

##   
## Call:  
## glm(formula = candy$chocolate ~ . - competitorname, family = "binomial",   
## data = candy)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.72224 -0.17612 -0.02787 0.01954 2.57898   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -10.29370 4.12040 -2.498 0.01248 \*   
## fruity -6.75305 2.20462 -3.063 0.00219 \*\*  
## caramel -1.85093 1.66750 -1.110 0.26700   
## peanutyalmondy -4.11907 2.98175 -1.381 0.16715   
## nougat -16.74818 3520.13323 -0.005 0.99620   
## crispedricewafer 14.98331 4725.35051 0.003 0.99747   
## hard 1.83504 1.80742 1.015 0.30997   
## bar 19.06799 3520.13379 0.005 0.99568   
## pluribus 0.22804 1.45457 0.157 0.87542   
## sugarpercent 0.12168 2.07707 0.059 0.95329   
## pricepercent 1.76626 2.24816 0.786 0.43208   
## winpercent 0.23019 0.08593 2.679 0.00739 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 116.407 on 84 degrees of freedom  
## Residual deviance: 25.802 on 73 degrees of freedom  
## AIC: 49.802  
##   
## Number of Fisher Scoring iterations: 19

Null deviance shows how well the response variable is predicted by a model that includes only the intercept

Residual deviance shows how well the response variable is predicted with inclusion of independent variables

#Predict the values and also to check the accuracy of the model.  
res <- predict(lgmodel,candy,type = "response")  
#confusion matrix  
table(ActualValue = candy$chocolate, PredictedValue = res > 0.5)

## PredictedValue  
## ActualValue FALSE TRUE  
## 0 47 1  
## 1 2 35

Confusion matrix: actual value is values from data set and predicted values is from the predicted values.

#accuracy  
(47+35)/(47+1+2+35)

## [1] 0.9647059

To calculate accuracy which is correct number of instances over total number of instances. This give the accuracy

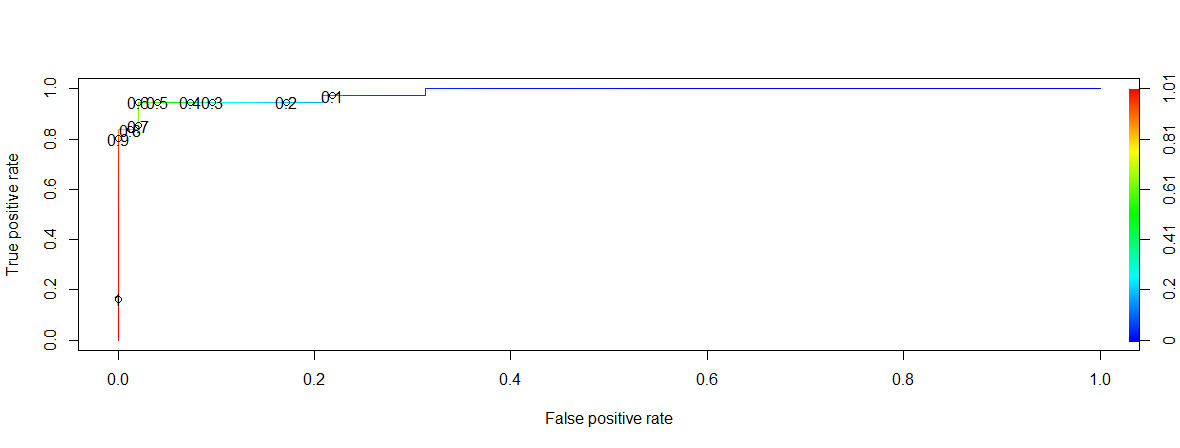
require(ROCR)

## Loading required package: ROCR

#define the ROCRPrediction and ROCRPerformance variables, this is used to plot a graph.  
acuracyPD = prediction(res,candy$chocolate)  
acuracyPF <- performance(acuracyPD,"tpr","fpr")

Tpr - true positive rate

Fpr - false positive rate  
#plotting the graph  
plot(acuracyPF, colorize = TRUE,print.cutoffs.at = seq(0.1,by=0.1))



This above graph shows most of the values have less False Positive rate and high true positive rate. This explains the accuracy of the predictions.

**Reason for selecting these algorithms:**

Is this a Fruit or a Vegetable? These kinds of questions are answered by Classification Algorithm.

Is this weird? These questions are with patterns, and answer by Anomaly Detection Algorithm.

How much or how many? These are quantifiable these are tackled by Regression algorithms.

How is this organized? This is dealt with clustering Alogorithms

What should I do Next? these types of questions are answer in Reinforcement Learning.

So, my research question is “Finding the best candy during Halloween”. To do that I need to consider few tasks, how many people prefer that candy over other? How much will be the cost comparison to other candies?

To get the answer to my question as I mentioned above to get the quantifiable results, we need to follow regression algorithms. That is the reason of choosing both linear and logistic regression.

**What is Regression**: this is drawing relationship between attributes.

There are three types of regression linear, logistic and Polynomial Regression. From these types I have taken Linear and Logistic regression.

**Linear regression:** when there is a linear relationship between independent and dependent variables. Using the linear regression algorithm, I have found the chocolates have 99% confident. So, if we use this attribute, we can get accurate results.

**Logistic regression:** When the dependent variable is categorical (0/1, True/False, Yes/No, Positive/Negative) in nature. Using this logistic regression and with help of the attribute with most impact on the dataset, finding its accuracy. Whether that attribute is a reason for the profit.

References:

1. <https://www.youtube.com/watch?v=Z5WKQr4H4Xk>
2. <http://www.learnbymarketing.com/tutorials/explaining-the-lm-summary-in-r/>
3. <http://rstudio-pubs-static.s3.amazonaws.com/491476_d4246618fe234dcfabb4284dd0f31fcf.html>