LogisticRegression.R

David

Fri Mar 09 19:36:07 2018

# 3/2018 - UofT DSSC Challenge  
# Logistic Regression + Data Vizualization  
  
# Import Libraries, setwd  
setwd("C:/Users/David/Desktop/UofT DSSC/Report")  
library("ggplot2")

## Warning: package 'ggplot2' was built under R version 3.4.2

library(ggmap)

## Warning: package 'ggmap' was built under R version 3.4.3

library(maps)

## Warning: package 'maps' was built under R version 3.4.3

library(mapdata)

## Warning: package 'mapdata' was built under R version 3.4.3

fuelmetrics <- read.csv("FuelStationsMetrics.csv")  
fuelmetrics <-subset(fuelmetrics,State=="Ontario" & PopularityScore > 2)  
dat <- read.csv("experiment\_data.csv")  
dat2<- subset(dat, AvgVol>1000)  
  
pop\_score <- dat$PopularityScore  
pop\_factor <- vector()  
  
# Classify Data - popularity scores > 2 and popularity scores <= 2  
for (i in seq(1,length(dat$PopularityScore))){  
 if (pop\_score[i] > 2){  
 pop\_factor[i] <- 1  
 }  
 else{  
 pop\_factor[i] <- 0  
 }  
}  
  
dat$PopularityScore <- pop\_factor   
dat$PopularityScore <- as.factor(dat$PopularityScore)  
  
### Create logistic regression model  
# variables  
# PopularityScore (1 if PopularityScore greater than 1, 0 otherwise)  
# AvgVol - average car valume around, mindistfuel - minimum distance to other fuelstation  
# rad1fuel - number of fuel stations close by, HasDiesel - whether shop has diesel  
  
# Clear from model A that popular gas stations are clustered together  
fita <- glm(PopularityScore~AvgVol+mindistfuel+rad1fuel+HasDiesel,family=binomial,data=dat)  
summary(fita)

##   
## Call:  
## glm(formula = PopularityScore ~ AvgVol + mindistfuel + rad1fuel +   
## HasDiesel, family = binomial, data = dat)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.7308 -0.4455 -0.3671 -0.1209 3.2047   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.588e+00 4.600e-01 -12.147 < 2e-16 \*\*\*  
## AvgVol 6.993e-04 7.212e-05 9.696 < 2e-16 \*\*\*  
## mindistfuel -1.041e+01 4.892e+00 -2.128 0.03334 \*   
## rad1fuel 7.760e-02 2.978e-02 2.605 0.00918 \*\*   
## HasDiesel 2.699e+00 4.558e-01 5.920 3.21e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1741.0 on 2951 degrees of freedom  
## Residual deviance: 1478.3 on 2947 degrees of freedom  
## AIC: 1488.3  
##   
## Number of Fisher Scoring iterations: 7

# For those looking at building a new gas station may make more sense to consider  
# just AvgVol and HasDiesel when selecting a site  
fitb <- glm(PopularityScore~AvgVol+HasDiesel,family=binomial,data=dat)  
summary(fitb)

##   
## Call:  
## glm(formula = PopularityScore ~ AvgVol + HasDiesel, family = binomial,   
## data = dat)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.8150 -0.4366 -0.3719 -0.1208 3.2643   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.581e+00 4.562e-01 -12.234 < 2e-16 \*\*\*  
## AvgVol 7.918e-04 6.945e-05 11.401 < 2e-16 \*\*\*  
## HasDiesel 2.683e+00 4.561e-01 5.884 4.02e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1741.0 on 2951 degrees of freedom  
## Residual deviance: 1498.9 on 2949 degrees of freedom  
## AIC: 1504.9  
##   
## Number of Fisher Scoring iterations: 7

# Assessing impact of AvgVol and has diesel on probability of being popular  
  
# Having diesel fuel increases probability of being Popular by  
coeff <- coefficients(fitb)  
pi\_HasDiesel <- exp(coeff['HasDiesel'])/(1+exp(coeff['HasDiesel']))  
pi\_HasDiesel

## HasDiesel   
## 0.9360287

# Analysis showing the gas stations that are the most popular have diesel by a wide margin  
xtabs(~HasDiesel+PopularityScore,data=dat)

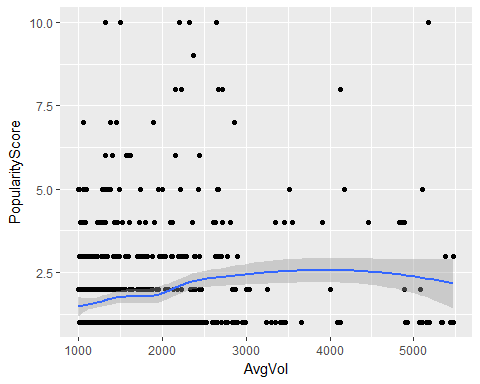
## PopularityScore  
## HasDiesel 0 1  
## 0 689 5  
## 1 2007 251

# Increase in AvgVol by to 2000 from 1000 cars increases makes probability of being Popular by  
pi\_AvgVol1000 <- (exp(2000\*coeff['AvgVol'])/(1+exp(2000\*coeff['AvgVol'])) - exp(1000\*coeff['AvgVol'])/(1+exp(1000\*coeff['AvgVol'])))  
pi\_AvgVol1000

## AvgVol   
## 0.1414939

### Vizualizations  
  
# Graph of relationship between avgvol and popularity scores  
ggplot(dat2,aes(x=AvgVol,y=PopularityScore)) + geom\_point() + geom\_smooth()

## `geom\_smooth()` using method = 'loess'



# Map of all Gas Stations by Popularity Score  
sbbox <- make\_bbox(lon = fuelmetrics$Longitude, lat = fuelmetrics$Latitude, f = .1)  
sq\_map <- get\_map(location = sbbox, maptype = "roadmap", source = "google")

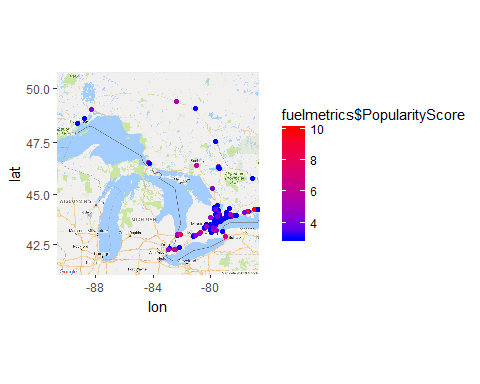
## Warning: bounding box given to google - spatial extent only approximate.

## converting bounding box to center/zoom specification. (experimental)

## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=46.01286,-83.66407&zoom=6&size=640x640&scale=2&maptype=roadmap&language=en-EN&sensor=false

ggmap(sq\_map) + geom\_point(data = fuelmetrics, mapping = aes(x = fuelmetrics$Longitude,  
 y = fuelmetrics$Latitude, color=fuelmetrics$PopularityScore)) + scale\_colour\_gradient(low="blue",high="red")

## Warning: Removed 27 rows containing missing values (geom\_point).



# Map of specific area of Toronto for analysis  
fuelmetrics <- subset(fuelmetrics, Geohash != 'dpz2sn2')  
  
sbbox <- make\_bbox(lon = c(-79.65,-79.60,-79.55), lat = c(43.65,43.70,43.75), f = .1)  
sq\_map <- get\_map(location = sbbox, maptype = "roadmap", source = "google")

## Warning: bounding box given to google - spatial extent only approximate.

## converting bounding box to center/zoom specification. (experimental)

## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=43.7,-79.6&zoom=13&size=640x640&scale=2&maptype=roadmap&language=en-EN&sensor=false

ggmap(sq\_map) + geom\_point(data = fuelmetrics, mapping = aes(x = fuelmetrics$Longitude,   
 y = fuelmetrics$Latitude, color=fuelmetrics$PopularityScore)) + scale\_colour\_gradient(low="blue",high="red")

## Warning: Removed 245 rows containing missing values (geom\_point).

