Bike Buyer

Jingning Zheng

Table of Contents:

I. Introduction

The question of this study is what factors generally influence the purchase of bicycles. The subjects of the research are mainly adults, mainly to understand some of the factors that influence people to buy bicycles.

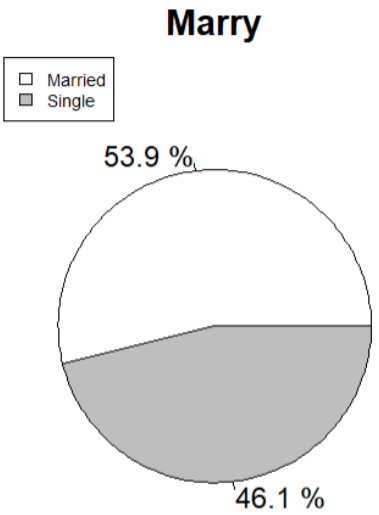
II. Description of Data

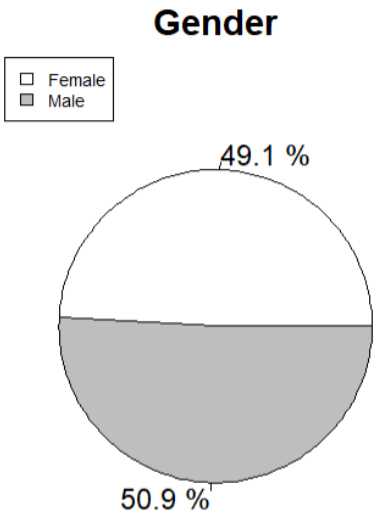
The data comes from Kaggle, by collecting the background information of 1000 people and whether to buy a bicycle. The background information collected includes marital status, gender, age, number of children, number of vehicles, whether they own a house, etc.

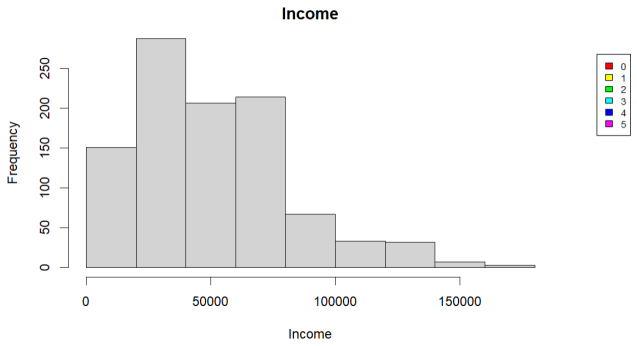
III. Preprocess of Data

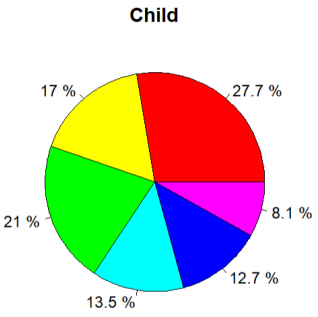
First, we check the data to exclude rows containing NA values or null values. Then copy the response variable to form a new column and convert it to 1 and 0. The response variable of this data is whether to buy a bicycle. Finally delete the ID and response variables that were copied.

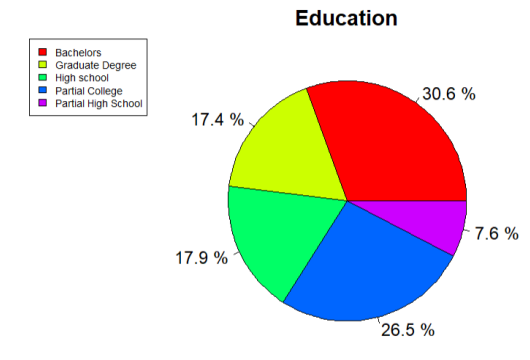
IV. Data Visualization

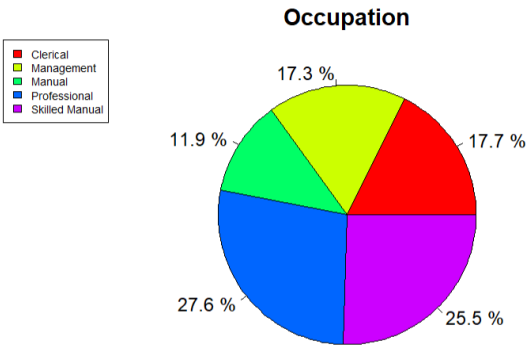


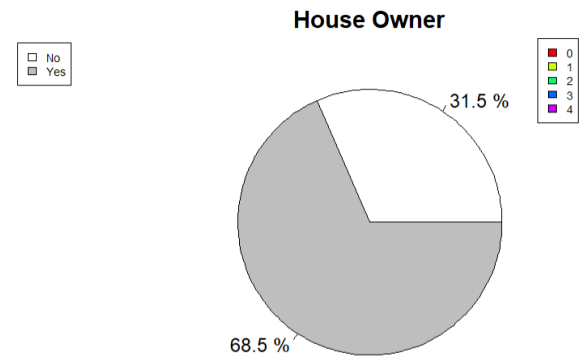


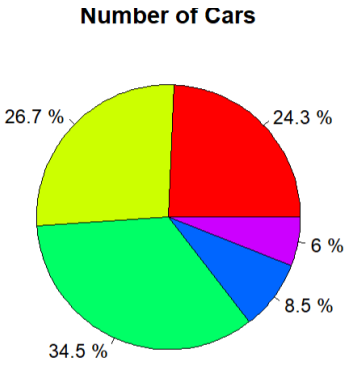


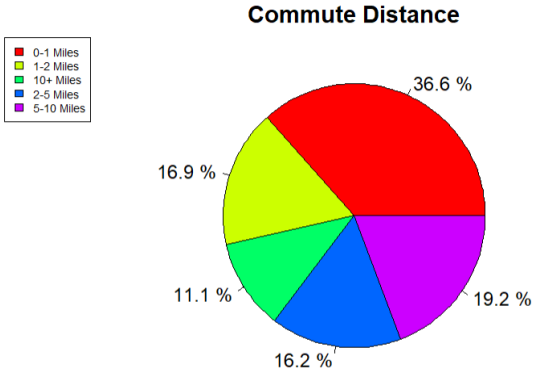


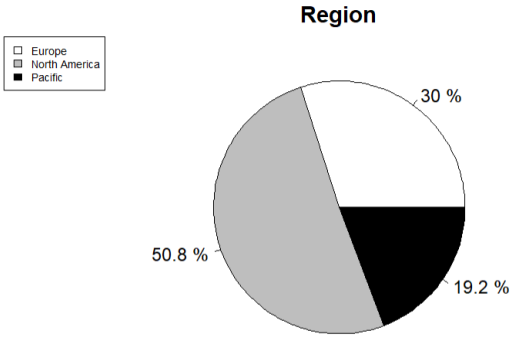


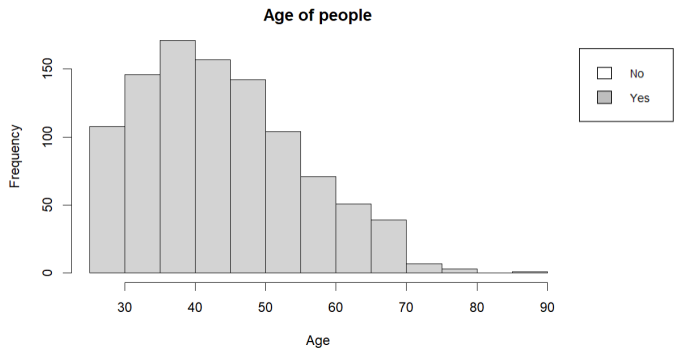


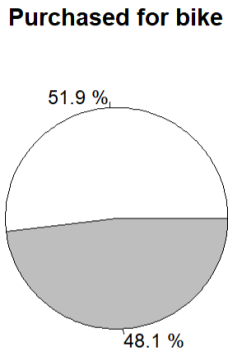












V. Model

Four models are fitted here, they are null model, full model, step model and select model. Among them, the step model uses both sides stepwise, and the select model is by deleting the non-significant variable in the full model.

VI. Model Analysis

A. Compare with other model

The selected model is the select model, and it is compared with the other three models by the Likelihood Ratio test. From the results, the model is better than the other three models.

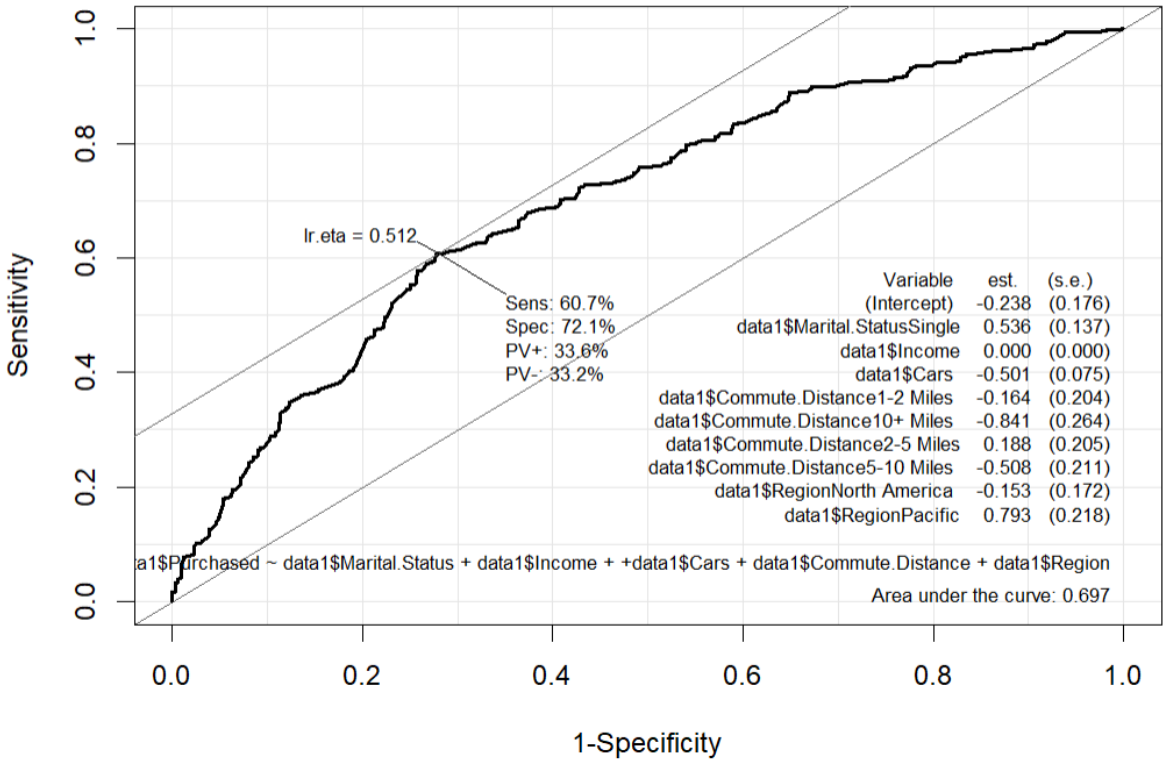
B. Leave one out cross-validation

From the verification results, after 1000 repeated tests, the accuracy of the model is 0.64, which is 64%. This means that the model is applied to the current data, and its accuracy is 64%.

C. K-fold cross validation

From the perspective of random folding, the select model has the highest accuracy among the four models. Then four k-fold cross validations are performed on the select model, and k is set to 10. The results obtained are all around 0.198.

D. Roc curve



From the roc curve graph, Sensitivity = 60.7% means that the ratio of observed number of people who purchased a bike to predicted number of people who purchased a bike is 0.607. Specificity = 72.1% means that the ratio of the observed number of people who have not purchased a bike to the predicted number of people who have not purchased a bike is 0.721. AUC=0.697 means that The applicability of the model to the current data is 69.7%

E. Correlation

From the perspective of correlation, the correlations of full model, step model and select model are all lower than 4.0, which indicates that the positive correlation between variables is relatively weak, and the correlation of select model is the lowest among the three models.

VII. Conclusion

From the results, the select model is more in line with the data, and the accuracy of the model is medium. For this data, we can know that for people to buy bicycles, marital status, whether they have a car, communication distance and region are the main factors affecting them, among which marital status is a positive influence, whether they have a car and communication distance are negative influences.

VIII. Appendix

```
library(MASS)

library(VGAM)

library(vcd)

library(pROC)

library(dplyr)

library(Epi)

library(lattice)

library(DAAG)

library(boot)



#Data

data1=read.csv("bike_buyers_clean.csv",header=T,na.strings = "")

data1=data1[complete.cases(data1), ]

row.has.na <- apply(data1, 1, function(x){any(is.na(x))})

sum(row.has.na)

data1 <- data1[!row.has.na,]

names(data1)

summary(data1)

data1$Purchased=data1$Purchased.Bike

data1$Purchased=ifelse(data1$Purchased=="Yes",1,0)

data1=data1[,-c(1,13)]



Marry = table(data1$Marital.Status)
```

```
piepercent<-paste(round(100*Marry/sum(Marry),2),"%")

pie(Marry,labels=piepercent,main="Marry",col=c("white","gray"))

legend("topleft",legend=c("Married","Single"),cex=0.6,fill=c("white","gray"))



Gender = table(data1$Gender)

piepercent<-paste(round(100*Gender/sum(Gender),2),"%")

pie(Gender,labels=piepercent,main="Gender",col=c("white","gray"))

legend("topleft",legend=c("Female","Male"),cex=0.6,fill=c("white","gray"))



hist(data1$Income,xlab = "Income",main = "Income")



Child = table(data1$Children)

piepercent<-paste(round(100*Child/sum(Child),2),"%")

pie(Child,labels=piepercent,main="Child",col=rainbow(length(Child)))

legend("topleft",legend=c("0","1","2","3","4","5"),

      cex=0.6,fill=rainbow(length(Child)))



Edu = table(data1$Education)

piepercent<-paste(round(100*Edu/sum(Edu),2),"%")

pie(Edu,labels=piepercent,main="Education",col=rainbow(length(Edu)))

legend("topleft",legend=c("Bachelors","Graduate Degree","High school","Partial
College","Partial High School"),

      cex=0.6,fill=rainbow(length(Edu)))
```

```
occ=table(data1$Occupation)
piepercent<-paste(round(100*occ/sum(occ),2),"%")
pie(occ,labels=piepercent,main="Occupation",col=rainbow(length(occ)))
legend("topleft",legend=c("Clerical","Management","Manual","Professional","Skilled Manual"),
       cex=0.6,fill=rainbow(length(occ)))

house = table(data1$Home.Owner)
piepercent<-paste(round(100*house/sum(house),2),"%")
pie(house,labels=piepercent,main="House Owner",col=c("white","gray"))
legend("topleft",legend=c("No","Yes"),cex=0.6,fill=c("white","gray"))


n.car=table(data1$Cars)
piepercent<-paste(round(100*n.car/sum(n.car),2),"%")
pie(n.car,labels=piepercent,main="Number of Cars",col=rainbow(length(n.car)))
legend("topleft",legend=c("0","1","2","3","4"),
       cex=0.6,fill=rainbow(length(n.car)))


distance=table(data1$Commute.Distance)
piepercent<-paste(round(100*distance/sum(distance),2),"%")
pie(distance,labels=piepercent,main="Commute Distance",col=rainbow(length(distance)))
legend("topleft",legend=c("0-1 Miles","1-2 Miles","10+ Miles","2-5 Miles","5-10 Miles"),
       cex=0.6,fill=rainbow(length(distance)))


region=table(data1$Region)
piepercent<-paste(round(100*region/sum(region),2),"%")
```

```
pie(region,labels=piepercent,main="Region",col=c("white","gray","black"))
legend("topleft",legend=c("Europe","North America","Pacific"),
       cex=0.6,fill=c("white","gray","black"))



hist(data1$Age,xlab = "Age",main = "Age of people")

pay=table(data1$Purchased)
piepercent<-paste(round(100*pay/sum(pay),2),"%")
pie(pay,labels=piepercent,main="Purchased for bike",col=c("white","gray"))
legend("topleft",legend=c("No","Yes"),cex=0.6,fill=c("white","gray"))



#### null model
mod0=glm(Purchased~1,data=data1,family = binomial(link="logit"))
summary(mod0)

#### full model

modf=glm(Purchased~.,data=data1,family = binomial(link="logit"))
summary(modf)

#### step model
mod1=step(modf,direction = "both",trace = F)
summary(mod1)
```

```
### select model
mod2=glm(Purchased~Marital.Status+Income+Children+Cars+Commute.Distance+Region,
        data=data1,family = binomial(link="logit"))
summary(mod2)


#### comparsion
anova(mod0, mod2, test="LRT")
anova(modf, mod2, test="LRT")
anova(mod1, mod2, test="LRT")


#leave one out cross-validation
prop <- sum(data1$Purchased)/nrow(data1)
prop
predicted <- as.numeric(fitted(mod1) > prop)
table1=xtabs(~ data1$Purchased + predicted)
table1
acc = (table1[1,1]+table1[2,2])/sum(table1)
acc
pihat <- vector(length=1000)
for (i in 1:1000) {
  pihat[i] <-
    predict(update(mod2, subset=-i),
            newdata=data1[i,], type="response")
}


yy <- as.numeric(data1$Purchased > 0)
yhat <- as.numeric(pihat >prop)
confusion <- table(yy, yhat)
```

```
confusion


acc = (confusion[1,1]+confusion[2,2])/sum(confusion)
acc


#### K-fold cross validation
cv.binary(mod0)
cv.binary(modf)
cv.binary(mod1)
cv.binary(mod2)
cost<-function(r,pi=0) mean(abs(r-pi)>0.6)
out0=cv.glm(data1,mod2,cost,K=10)
names(out0)
out1=cv.glm(data1,mod2,cost,K=10)
out2=cv.glm(data1,mod2,cost,K=10)
out3=cv.glm(data1,mod2,cost,K=10)


out0$delta
out1$delta
out2$delta
out3$delta



#### roc curve for each model


### null model
rocplot0 <- roc(Purchased ~ fitted(mod0), data=data1)
plot.roc(rocplot0, legacy.axes=TRUE)
```

```
auc(rocplot0)


### full model
rocplot1 <- roc(Purchased ~ fitted(modf), data=data1)
plot.roc(rocplot1, legacy.axes=TRUE)
auc(rocplot1)


### step model
rocplot2 <- roc(Purchased ~ fitted(mod1), data=data1)
plot.roc(rocplot2, legacy.axes=TRUE)
auc(rocplot2)


### select model
rocplot2 <- roc(Purchased ~ fitted(mod2), data=data1)
plot.roc(rocplot2, legacy.axes=TRUE)
auc(rocplot2)


ROC(form=data1$Purchased~data1$Marital.Status+data1$Income+
    +data1$Cars+data1$Commute.Distance+data1$Region,plot="ROC")


#### correlation
cor(data1$Purchased, fitted(modf))
cor(data1$Purchased, fitted(mod1))
cor(data1$Purchased, fitted(mod2))
```