Bike Buyer

Jingning Zheng

Justin Fang

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.

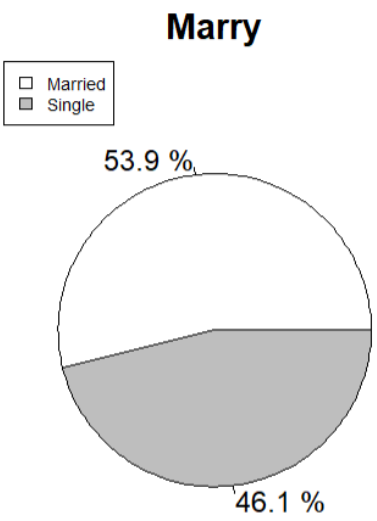
II. Description of Data

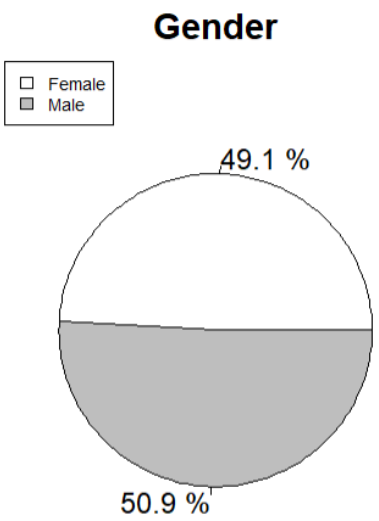
The data comes from Kaggle, by collecting the background information of 1000 people and whether to buy a bicycle. The background information collected includes marital status, gender, age, number of children, number of vehicles, whether they own a house, etc.

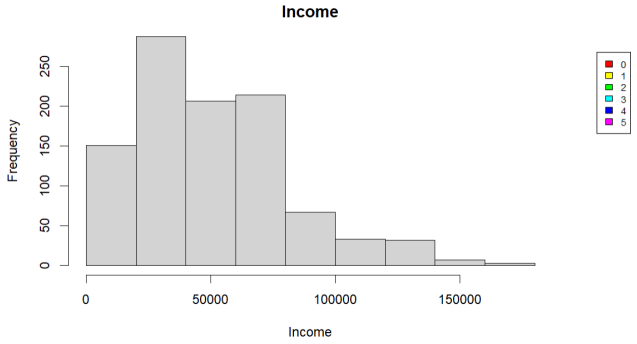
III. Preprocess of Data

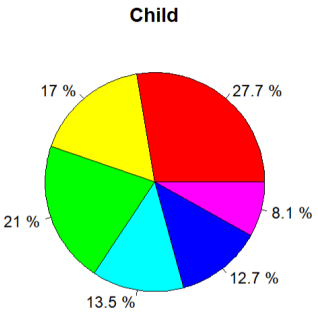
First, we check the data to exclude rows containing NA values or null values. Then copy the response variable to form a new column and convert it to 1 and 0. The response variable of this data is whether to buy a bicycle. Finally delete the ID and response variables that were copied.

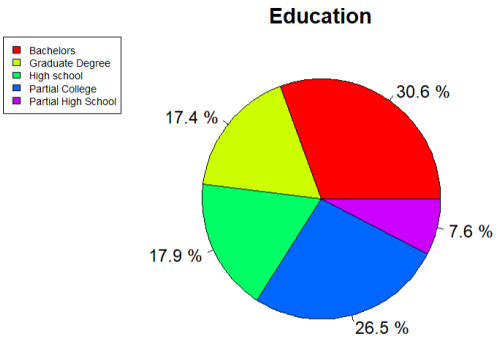
IV. Data Visualization

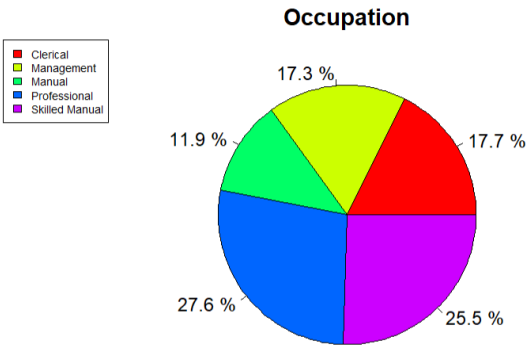

Marry
Married
Single
53.9 %
46.1 %

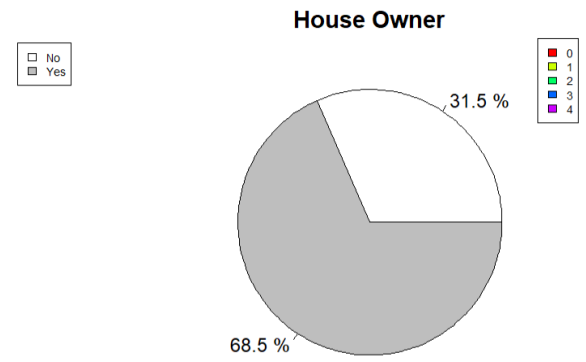

Gender
Female
Male
49.1 %
50.9 %

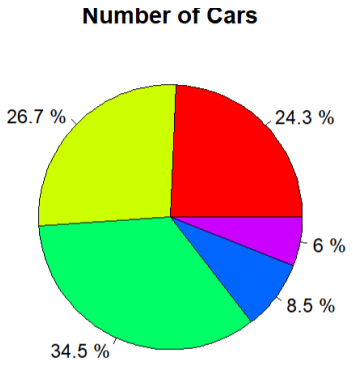

Income
Frequency
0
50
100
150
200
250
0
50000
100000
150000
Income
0
1
2
3
4
5

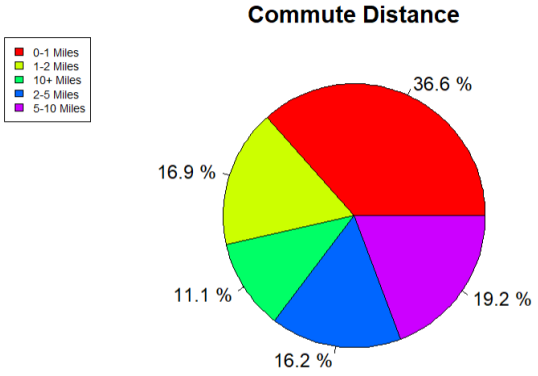

Child
17 %
27.7 %
21 %
8.1 %
12.7 %
13.5 %

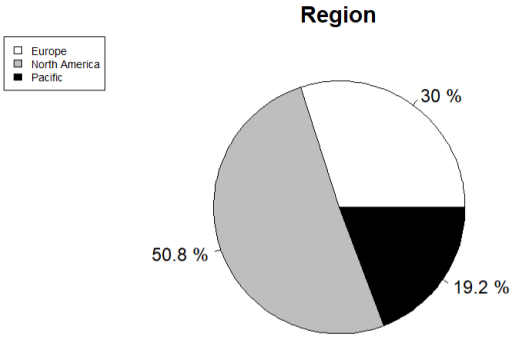

Education
Bachelors
Graduate Degree
High school
Partial College
Partial High School
30.6 %
17.4 %
7.6 %
17.9 %
26.5 %

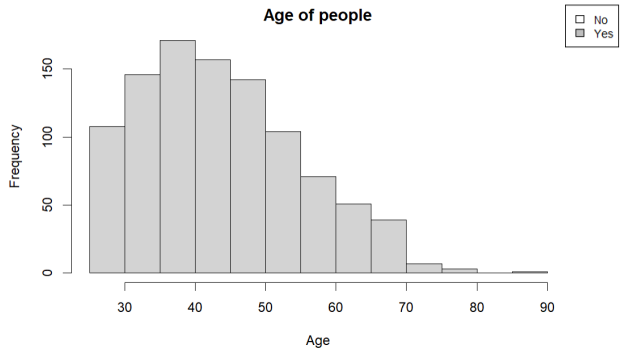

Occupation
Clerical
Management
Manual
Professional
Skilled Manual
17.3 %
11.9 %
17.7 %
27.6 %
25.5 %

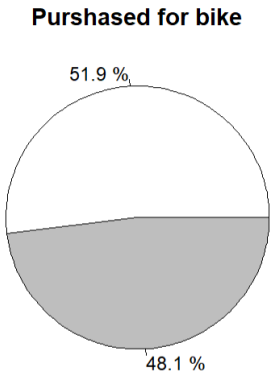












V. Model

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 mode

```
anova(mod0, mod2, test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Purshased ~ 1
## Model 2: Purshased ~ Marital.Status + Income + Children + Home.Owner +
##     Cars + Commute.Distance + Region
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       999     1384.8
## 2       988     1256.6 11   128.25 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(modf, mod2, test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Purshased ~ Marital.Status + Gender + Income + Children + Education +
##     Occupation + Home.Owner + Cars + Commute.Distance + Region +
##     Age
## Model 2: Purshased ~ Marital.Status + Income + Children + Home.Owner +
##     Cars + Commute.Distance + Region
##   Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
## 1       978     1239.0
## 2       988     1256.6 -10  -17.652  0.06112 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod1, mod2, test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Purshased ~ Marital.Status + Income + Children + Occupation +
##     Home.Owner + Cars + Commute.Distance + Region
## Model 2: Purshased ~ Marital.Status + Income + Children + Home.Owner +
##     Cars + Commute.Distance + Region
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       984     1247.0
## 2       988     1256.6 -4   -9.572  0.04829 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The first comparison between the null model (mod0) and the reduced model with multiple predictors (mod2) shows that the latter significantly

improves the fit ($p < 2.2e-16$). The second comparison between the full model (modf) and a reduced model (mod2) suggests that removing certain predictors from the full model does not significantly affect the fit ($p = 0.06112$), indicating that those predictors may not contribute significantly. However, the third comparison between a stepwise model (mod1) and the reduced model (mod2) reveals that the stepwise model is significantly better ($p = 0.04829$), suggesting that the predictors included in the stepwise model contribute meaningfully to the prediction of the outcome variable. In conclusion to this analysis, it demonstrates that including predictors generally improves the model's fit compared to the null model and that the selected predictors in the stepwise model are valuable for prediction.

B. Leave one out cross-validation

```
##                 predicted
## data1$Purshased   0   1
##               0 338 181
##               1 167 314
```

```
##    yhat
## yy    0   1
##   0 329 190
##   1 165 316
```

Table1(left) shows the comparison between the actual purchased values and the predicted values. The accuracy measure was calculated and was found to be 65.2%. The confusion matrix values indicate that there were 329 true positives, 190 false negatives, 165 false positives, and 316 true negatives. The accuracy was recalculated using the confusion matrix, yielding a value of 64.5%. Overall, this analysis demonstrates the effectiveness of the logistic regression model in predicting purchases, with a proportion-based prediction approach and leave-one-out cross-validation providing a 65.2% accuracy.

C. K-fold cross validation

```
cv.binary(mod0)
```

```
##
## Fold:  3 6 9 2 4 8 10 5 1 7
## Internal estimate of accuracy = 0.519
## Cross-validation estimate of accuracy = 0.519
```

```
cv.binary(modf)
```

```
##
## Fold:  9 4 6 7 2 3 5 10 8 1
## Internal estimate of accuracy = 0.655
## Cross-validation estimate of accuracy = 0.636
```

```
cv.binary(mod1)
```

```
##
## Fold:  7 9 8 6 3 2 10 1 5 4
## Internal estimate of accuracy = 0.657
## Cross-validation estimate of accuracy = 0.622
```

```
cv.binary(mod2)
```

```
##
## Fold:  7 2 8 4 9 1 3 10 5 6
## Internal estimate of accuracy = 0.67
## Cross-validation estimate of accuracy = 0.653
```

For the null model (mod0), the internal estimate of accuracy and the cross-validation estimate of accuracy are both approximately 51.9%. This suggests that the model performs consistently across different folds, as the internal and cross-validation estimates are similar.

For the full model (modf), the internal estimate of accuracy is approximately 65.5%, while the cross-validation estimate of accuracy is slightly lower at 63.6%. This indicates that the model may exhibit some variability in performance across different folds, with a slightly lower accuracy when evaluated through cross-validation.

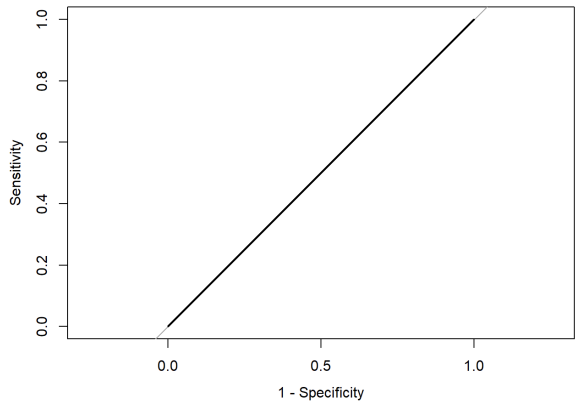
For the step model (mod1), the internal estimate of accuracy is approximately

65.7%, while the cross-validation estimate of accuracy is slightly lower at 62.2%. Similar to modf, this suggests some variability in performance across folds and a slightly lower accuracy when evaluated through cross-validation.

For reduced mode (mod2), the internal estimate of accuracy is approximately 67%, and the cross-validation estimate of accuracy is slightly lower at 65.3%. This indicates that the model performs consistently across folds, with a relatively high accuracy .

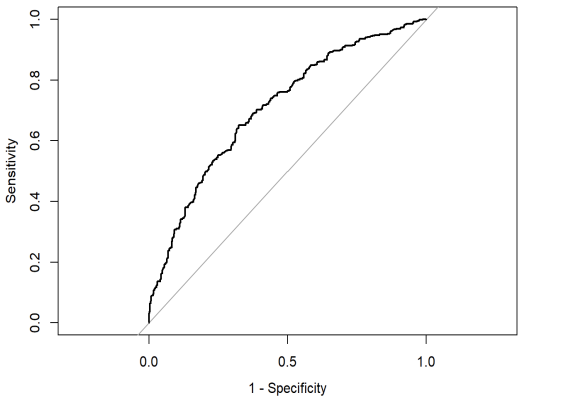
The cv.glm function is then used for cross-validation on the dataset, which includes background information such as marital status, gender, age, number of children, number of vehicles, and homeownership. For mod0: delta = 0.0170, 0.0176; For mod1: delta = 0.016, 0.017; For mod2: delta = 0.0170, 0.0172. We can see that for all the models, the differences between the internal and cross-validation estimates of accuracy are relatively small, indicating reasonable generalization performance. The variations in the delta values between the models are also not significant. Therefore we can come to a suggestion that the cross-validation results demonstrate the models are performing reasonably well and have the potential to make accurate predictions on unseen data.

D. Roc curve



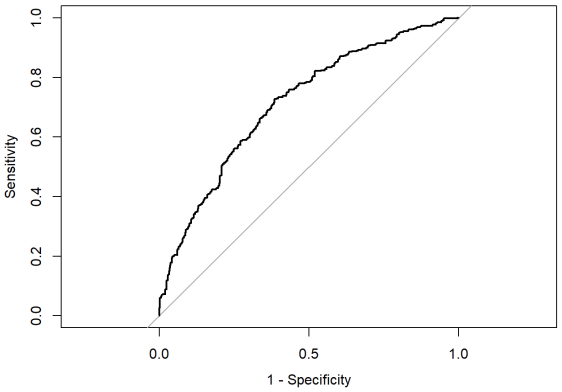

```
auc(rocplot0)
```

```
## Area under the curve: 0.5
```



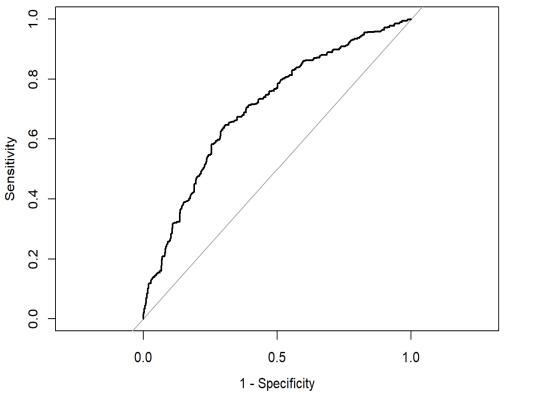

```
auc(rocplot2)
```

```
## Area under the curve: 0.7076
```




```
auc(rocplot1)
```

```
## Area under the curve: 0.7132
```




```
auc(rocplot2)
```

```
## Area under the curve: 0.7031
```

The area under roc curve for the null model is 0.5 (upper left); the area under roc curve for the full model is 0.7076 (upper right); The area under roc curve for the the step model is 0.7132 (lower left); lastly the area under roc curve for the null model is 0.7031 (lower right). The area under roc curve values for all the models are greater than 0.5, indicating some discriminatory power. The full model, stepwise model, and final model all exhibit similar values, indicating reasonably good predictive performance. However, the area under roc curve values are not extremely high, suggesting that there is still room for improvement in the models. Therefore, the analysis suggests that the selected predictor variables have some predictive value for the purchased variable, but there may be other

factors not considered in the models that could further enhance the prediction accuracy.

E. Correlation

```
#### correlation
cor(data1$Purshased, fitted(modf))
```

```
## [1] 0.3707286
```

```
cor(data1$Purshased, fitted(mod1))
```

```
## [1] 0.3605808
```

```
cor(data1$Purshased, fitted(mod2))
```

```
## [1] 0.3513227
```

The correlation analysis indicates the correlation between the actual bicycle purchases (Purchased) and the fitted values from the 4 models. The correlation coefficient quantifies the linear relationship between these variables. For the full model (modf), the correlation coefficient is 0.3707, for the step model (mod1) it is 0.3606, and for the final reduced model (mod2) it is 0.3513. These correlation coefficients indicate a moderate positive linear relationship between the predicted probabilities of purchase from the models and the actual purchase outcomes. In this case, the positive correlations suggest that as the predicted probabilities of purchase increase, the likelihood of actual purchases also tends to increase.

VII. Conclusion

In this study, we investigated the factors that influence the purchase of bicycles among adults. By analyzing data collected from 1000 individuals, we conducted various analyses and model comparisons to understand the significance of different predictors in predicting bicycle purchases.

We process and analyze the data into different models and make different model analyses that reveal predictors significantly improved the fit of the model compared to the null model. The stepwise model, in particular, demonstrated better performance and suggested that the predictors included in the model contribute meaningfully to predicting the outcome variable. This highlights the value of selected predictors in understanding and predicting bicycle purchases. Other model analysis including leave-one-out cross-validation approach provided an accuracy of 65.2%, indicating the effectiveness of the logistic regression model in predicting purchases. K-fold cross-validation further supported the consistent performance of the models across different folds, with relatively high accuracy rates. Moreover, the correlation analysis showed a moderate positive linear relationship between predicted probabilities of purchase and actual purchase outcomes, reinforcing the validity of the models and their predictive value.

In conclusion, we came to understand from this study about the factors influencing bicycle purchases among adults. The findings demonstrate the importance of considering various predictors for business and environmental improvement. If we were to conduct this project again, we would consider other external factors that may potentially influence bicycle purchase such as economic conditions, cultural attitudes towards cycling, availability of bike-sharing programs in locals.

VIII. Appendix

```
#### Jingning Zheng, Justin Fang

#### Data: Bike Buyer


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))
```