BUDT758T

DATA MINING AND PREDICTIVE ANALYTICS

Homework 2

NAME (in capitals): ___________AKHIL GUPTA____

- Please submit on Canvas.
- Your submission should consist of this document (with answers filled in in the appropriate places).
- Please ensure that answers are appropriately numbered and clearly legible.
- In the space below please enter the following text and initial below: “I pledge on my honor that I have not given or received unauthorized assistance on this assignment.”

HONOR PLEDGE: I pledge on my honor that I have not given or received unauthorized assistance on this assignment. YOUR INITIALS: AG

The goal of this homework is to introduce you to classification concepts. You will develop (1) a linear probability model and (2) a logistic regression model. You will need to create random partitions of a data set, build your model on the training data set and then compute prediction errors using the test data set.

The Assignment

The data in the accompanying file "VoterPref.csv" (posted on Canvas) contains data from a survey of random sample of registered voters in a state. The subjects were asked whether they were "For" or "Against" a proposal on the ballot to increase the state sales tax by 0.5%, with the stipulation that the additional tax revenues be spent on education. In addition to their position on the proposition, some additional demographic information is collected. The variables in the data set are:

PREFERENCE	"For" or "Against"
AGE	Years of age at time of survey
INCOME	Annual income in thousands of US dollars
GENDER	"M" or "F"

The intent of the survey is to develop a strategy to target individuals for a marketing campaign designed to "get out the vote".

(1) Data Preparation

```
getwd()
```

```
## [1] "D:/Sem2/1. self/dataMininingPredictiveAnalysis/HW/2. HW2"
```

```
setwd("D:/Sem2/1. self/dataMininingPredictiveAnalysis/HW/2. HW2")
getwd()
```

```
## [1] "D:/Sem2/1. self/dataMininingPredictiveAnalysis/HW/2. HW2"
```

a. Read the data set in *R*. For the PREFERENCE variable ensure that "Against" is the success class

```
pref<-read.csv('VoterPref.csv')
head(pref)
```

```
##   AGE INCOME GENDER PREFERENCE
## 1  16  39.06      F        For
## 2  36  68.83      F        For
## 3  50 113.20      F        For
## 4  33 122.76      M        For
## 5  26 107.49      M        For
## 6  42  86.95      M        For
```

```
L_PREF<-pref$PREFERENCE=='Against'
pref<-cbind(pref, L_PREF)
```

b. Set the seed to 123457

```
set.seed(123457)
```

c. Randomly partition the data set into the *training* and *test* data sets. The proportion of observations in the training data set should be 70%. The remaining 30% of observations should be in the test data set.

```
train_ind<-sample(seq_len(nrow(pref)),size= .7*nrow(pref))
train<-pref[train_ind, ]
test<-pref[-train_ind, ]
nrow(train)

## [1] 700

nrow(test)

## [1] 300
```

(2) Exploratory analysis of the *training* data set

a. Construct boxplots of INCOME and AGE (broken up by values of PREFERENCE). Present the plot as **Exhibit A**. What do you observe?

```
boxplot(INCOME~PREFERENCE,data=train, main="INCOME PREFERENCE", xlab="PREFERENCE
", ylab="INCOME")


boxplot(AGE~PREFERENCE,data=train, main="AGE PREFERENCE", xlab="PREFERENCE", yla
b="AGE")
```

Please refer to Exhibit A for the plots

The box plot shows that the people with lower income are more likely to vote “AGAINST” the proposition than with people with higher income who may vote “FOR” the proposition.

The people who are elder are more likely to vote “AGAINST” for the proposition while the younger people have a higher likeliness of voting “FOR” the proposition.

b. Construct a table for PREFERENCE showing proportions for and against.

```
train_PEFERENCE_tab <- table(train$PREFERENCE)
#2b

train_PEFERENCE_tab <- table(train$PREFERENCE)
train_PEFERENCE_tab

##
## Against     For
##     133     567

prop.table(train_PEFERENCE_tab)
```

```
##
## Against     For
##    0.19    0.81
```

c. Construct a two-way table for count of PREFERENCE broken up by GENDER (i.e. what are the numbers of men and women who are for and against the proposition).

```
train_PREFERENCE_GENDER_tab<- table(train$PREFERENCE, train$GENDER)
train_PREFERENCE_GENDER_tab

##
##               F   M
##   Against  71  62
##   For     266 301

prop.table(train_PREFERENCE_GENDER_tab)

##
##                      F          M
##   Against 0.10142857 0.08857143
##   For     0.38000000 0.43000000
```

(3) Run a linear regression model of PREFERENCE on the demographic variables. Use only the training data set for this.

```
fit<-lm(as.numeric(L_PREF)~AGE +INCOME+factor(GENDER), data = train)
summary(fit)

## Call:
## lm(formula = as.numeric(L_PREF) ~ AGE + INCOME + factor(GENDER),
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60250 -0.22397 -0.06213  0.16475  0.87091
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.3824368  0.0620009   6.168 1.17e-09 ***
## AGE              0.0197151  0.0014403  13.689  < 2e-16 ***
## INCOME          -0.0099003  0.0005948 -16.646  < 2e-16 ***
## factor(GENDER)M -0.0469981  0.0236309  -1.989   0.0471 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3119 on 696 degrees of freedom
## Multiple R-squared:  0.3714, Adjusted R-squared:  0.3687
## F-statistic: 137.1 on 3 and 696 DF,  p-value: < 2.2e-16
```

a. Compute the average error, RMSE, and MAPE for both in-sample predictions (i.e. for the training data set) and the out-of-sample predictions (i.e. for the test data set). Use predicted values from the regression equation (do not do the classification yet).

```
#3a
predicted_train<-predict(fit, newdata = train)
actual_train<-as.numeric(train$L_PREF)
Metrics<-c("AE","RMSE","MAE")
x1<-mean(actual_train-predicted_train)
x2<-sqrt(mean((actual_train-predicted_train)^2))
x3<-mean(abs(actual_train-predicted_train))

Values<- c(x1,x2,x3)
X_train<-data.frame(Metrics, Values)
X_train

##   Metrics        Values
## 1      AE -2.191780e-16
## 2    RMSE  3.110287e-01
## 3     MAE  2.476011e-01

#************************************#
predicted_test<-predict(fit, newdata = test)
actual_test<-as.numeric(test$L_PREF)
Metrics<-c("AE","RMSE","MAE")
x4<-mean(actual_test-predicted_test)
x5<-sqrt(mean((actual_test-predicted_test)^2))
x6<-mean(abs(actual_test-predicted_test))

Values<- c(x4,x5,x6)
X_test<-data.frame(Metrics, Values)
X_test

##   Metrics       Values
## 1      AE -0.006721264
## 2    RMSE  0.319082635
## 3     MAE  0.256544188
```

b. For which data set are these errors smaller?

For the training dataset the errors are smaller as we are training our dataset from the same dataset as we are predicting.

c. Use a cutoff of 0.5 and do the classification. Compute the confusion matrix for both in-sample and out-of-sample predictions.

```
#3c

t_train <- ifelse(predicted_train > 0.5,1,0)
confusion_m_train<-table(as.numeric(train$L_PREF),t_train)
confusion_m_train
```

```
##    t_train
##       0   1
##   0 557  10
##   1  75  58
```

Table 1

	Predicted	
Actual	1 (Against)	0 (For)
1 (Against)	58	75
0 (For)	10	557

```
confusion_m_train_probability<-confusion_m_train/sum(confusion_m_train)
confusion_m_train_probability

##    t_train
##              0          1
##   0 0.79571429 0.01428571
##   1 0.10714286 0.08285714
```

Table 2

	Predicted	
Actual	1 (Against)	0 (For)
1 (Against)	.083	.107
0 (For)	.014	.796

```
t_test <- ifelse(predicted_test > 0.5,1,0)
confusion_m_test<-table(as.numeric(test$L_PREF),t_test)
confusion_m_test

##    t_test
##       0   1
##   0 234   8
##   1  33  25
```

Table 3

	Predicted	
Actual	1 (Against)	0 (For)
1 (Against)	25	33
0 (For)	8	234

```
confusion_m_test_probability<-confusion_m_test/sum(confusion_m_test)
confusion_m_test_probability

##    t_test
##              0          1
##   0 0.78000000 0.02666667
##   1 0.11000000 0.08333333
```

Table 4

	Predicted	
Actual	1 (Against)	0 (For)
1 (Against)	.083	.110
0 (For)	.027	.780

d. Compare the two confusion matrices.

Here, I have compared the confusion matrix probabilities of both the training and test data.

```
confusion_m_train_probability-confusion_m_test_probability

##    t_train
##                0            1
##   0  0.0157142857 -0.0123809524
##   1 -0.0028571429 -0.0004761905
```

Table 5

	Predicted	
Actual	1 (Against)	0 (For)
1 (Against)	-.005	-.003
0 (For)	-.012	.016

- We observe that True positive cases are almost equal
- Also False negative cases are almost equal
- False positive probability of training set is .012 less than that of the test data
- True negative probability of training set is .016 more than that of the test data

Table 6: Makinh this table with Table 1 and Table 3

	TRAIN		TEST	
	Predicted		Predicted	
Actual	1 (Against)	0 (For)	1 (Against)	0(For)
1 (Against)	58	75	25	33
0 (For)	10	557	8	234

- We observe that in the training dataset the true positive cases (i.e. cases when the Actual is Against and Predicted is also against) is 58 out of 700 and in the test dataset the true positive cases are 25 out of 300
- We observe that in the training dataset the false negative cases (i.e. cases when the Actual is Against and Predicted is for) is 75 out of 700 and in the test dataset the true positive cases are 33 out of 300
- We observe that in the training dataset the false positive cases (i.e. cases when the Actual is For and Predicted is Against) is 10 out of 700 and in the test dataset the true positive cases are 8 out of 300
- We observe that in the training dataset the true negative cases (i.e. cases when the Actual is For and Predicted is also For) is 557 out of 700 and in the test dataset the true positive cases are 234 out of 300

(4) Run a logistic regression model of PREFERENCE on the demographic variables. Use only the training data set for this.

a. Present the output as **Exhibit B**.

```
#4a.
fit_logistic <- glm(as.numeric(L_PREF)~AGE+INCOME+factor(GENDER), family = "binom
ial", data = train)
summary(fit_logistic)
```

Please find the output attached as Exhibit B

b. Provide a precise interpretation of the coefficient of AGE.

b_{Age} =.23953
e^(b_{Age})= 1.27
We imply that with 1 year increase in age, the odds of voting against increase by a factor of 1.27, for those with same gender and income.

c. Provide a precise interpretation of the coefficient of the gender variable.

b_{GENDER}= -0.53005

e^(b_{GENDER})= 0.589

We imply that the odds of a **male** customer preferring to vote against the proposition are 0.589 times the odds of a **female** customer *of the same income and age* preferring to vote against the proposition.

d. Use a cutoff of 0.5 and do the classification. Compute the confusion matrix for both in-sample and out-of-sample predictions (using the training and test data sets respectively).

```
#4d.
predicted_train_logistic <- predict(fit_logistic, newdata = train, type = "respo
nse")

tlogistic_train <- ifelse(predicted_train_logistic > 0.5,1,0)

confusion_m_logistic_train<-table(as.numeric(train$L_PREF),tlogistic_train)

confusion_m_logistic_train_probability<-confusion_m_logistic_train/sum(confusion
_m_logistic_train)
confusion_m_logistic_train_probability

##    tlogistic_train
##              0          1
##   0 0.77857143 0.03142857
##   1 0.06714286 0.12285714

#*****************************************#
predicted_test_logistic <- predict(fit_logistic, newdata = test, type = "respons
e")
```

```r
tlogistic_test <- ifelse(predicted_test_logistic > 0.5,1,0)

confusion_m_logistic_test<-table(as.numeric(test$L_PREF),tlogistic_test)
confusion_m_logistic_test_probability<-confusion_m_logistic_test/sum(confusion_m
_logistic_test)
confusion_m_logistic_test_probability
```

```
##    tlogistic_test
##              0          1
##   0 0.75000000 0.05666667
##   1 0.08000000 0.11333333
```

e. Compare the two confusion matrices and compare with the corresponding matrices in question (3) above.

```r
#4e
compare_test<-confusion_m_logistic_train - confusion_m_train
compare_test

##    tlogistic_train
##       0   1
##   0 -12  12
##   1 -28  28
```

Table 7

	Predicted	
Actual	1 (Against)	0 (For)
1 (Against)	28	-28
0 (For)	12	-12

- We observe that in the **training** dataset for the **logistic** model the true positive cases (i.e. cases when the Actual is Against and Predicted is also Against) is 28 more than the **training** dataset for the **linear** model
- We observe that in the **training** dataset for the **logistic** model the false negative cases (i.e. cases when the Actual is Against and Predicted is For) is 28 less than the **training** dataset for the **linear** model
- We observe that in the **training** dataset for the **logistic** model the false positive cases (i.e. cases when the Actual is For and Predicted is Against) is 12 more than the **training** dataset for the **linear** model
- We observe that in the **training** dataset for the **logistic** model the true negative cases (i.e. cases when the Actual is For and Predicted is also For) is 12 less than the **training** dataset for the **linear** model

```r
compare_train<-confusion_m_logistic_test-confusion_m_test
compare_train

##    tlogistic_test
##      0  1
##   0 -9  9
##   1 -9  9
```

Table 8

	Predicted		
Actual		1 (Against)	0 (For)
	1 (Against)	9	-9
	0 (For)	9	-9

- We observe that in the **test** dataset for the **logistic** model the true positive cases (i.e. cases when the Actual is Against and Predicted is also Against) is 9 more than the **test** dataset for the **linear** model
- We observe that in the **test** dataset for the **logistic** model the false negative cases (i.e. cases when the Actual is Against and Predicted is For) is 9 less than the **test** dataset for the **linear** model
- We observe that in the **test** dataset for the **logistic** model the false positive cases (i.e. cases when the Actual is For and Predicted is Against) is 9 more than the **test** dataset for the **linear** model
- We observe that in the **test** dataset for the **logistic** model the true negative cases (i.e. cases when the Actual is For and Predicted is also For) is 9 less than the **test** dataset for the **linear** model

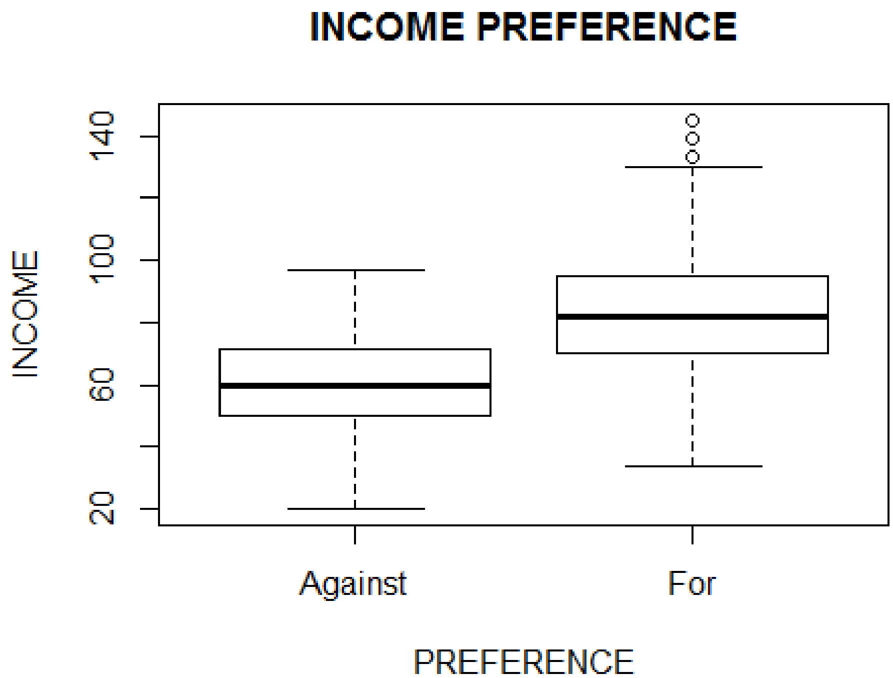
f. Compute the predicted probability for voting *against* the proposition for an individual who is a female, is 36 years old, and has an income $70,000.

```
#4f
spcase <- data.frame(AGE = 36,INCOME = 70,GENDER = "F")
predicted_prob <- predict(fit_logistic, newdata = spcase, type= "response")
predicted_prob

##         1
## 0.3838874

# Verifying 4f
#p(Against) = 1/(1+e(-logit))
#logit = 0.13300 + 0.23953(AGE) -0.13184(INCOME) - 0.53005(M=1,F=0)
logit <- 0.13300 + 0.23953*36 -0.13184*70 + 0
logit

## [1] -0.47272

x <- 1 + exp(-logit)
1/x

## [1] 0.3839727
```

Notes: Please see the additional slides on the *R* **table** function to generate confusion matrices. The *R* **ifelse** function will conveniently allow you to do the classification. Finally, the **predict** function will also work for the logistic case. Note however that it will give you the predicted logit. If you pass it an additional argument (type = "response") you will get predicted probabilities. E.g.

p <- predict(fit, newdata=df, type = "response")

Exhibit A


INCOME PREFERENCE
INCOME
20
60
100
140
Against
For
PREFERENCE

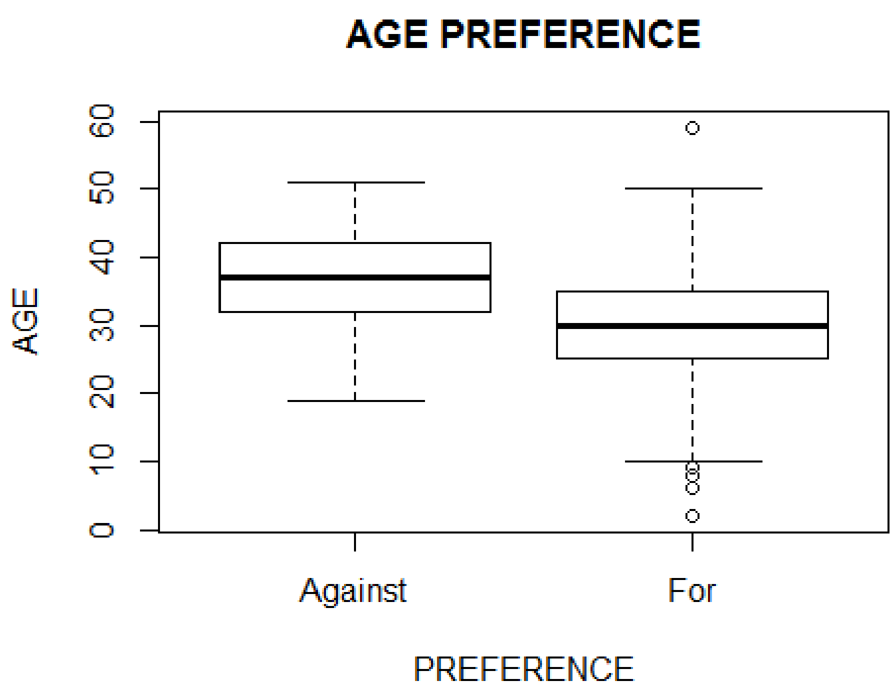

AGE PREFERENCE
AGE
0
10
20
30
40
50
60
Against
For
PREFERENCE

Exhibit B

```
##
## Call:
## glm(formula = as.numeric(L_PREF) ~ AGE + INCOME + factor(GENDER),
##     family = "binomial", data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.23799  -0.38579  -0.13440  -0.02922   2.81772
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.13300    0.76992   0.173    0.863
## AGE              0.23953    0.02462   9.729   <2e-16 ***
## INCOME          -0.13184    0.01268 -10.398   <2e-16 ***
## factor(GENDER)M -0.53005    0.27957  -1.896    0.058 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 680.71  on 699  degrees of freedom
## Residual deviance: 340.35  on 696  degrees of freedom
## AIC: 348.35
##
## Number of Fisher Scoring iterations: 7
```