****

**BUDT758T  
  
DATA MINING AND PREDICTIVE ANALYTICS**

**Individual Assignment 2**

**NAME (in capitals): ROHIT KOTHAVADE**

* Please see the instructions at <https://docs.google.com/document/d/1uwOFS-LVKDBAzjEonmfggJMAmNWhWxAJRbL71nsVg4A/edit?usp=sharing> and submit on Canvas.
* Your submission should consist of this document (with the answers filled 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: RVK |

This is an individual assignment. Your submission must represent your own work.

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. There are a couple of helpful hints at the end of the assignment.

**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
   1. Read the data set in *R*. For the PREFERENCE variable ensure that “Against” is the success class (i.e. the class with higher level – e.g. “1” for binary variable)

Ans – Here we can see that “Against” is the success class because I have changed the levels. For = 1,

Against = 2.

dfk$PREFERENCE<-factor(dfk$PREFERENCE,levels=c("For","Against"))  
str(dfk)

## 'data.frame': 1000 obs. of 4 variables:  
## $ AGE : int 16 36 50 33 26 42 29 17 27 39 ...  
## $ INCOME : num 39.1 68.8 113.2 122.8 107.5 ...  
## $ GENDER : Factor w/ 2 levels "F","M": 1 1 1 2 2 2 1 1 2 1 ...  
## $ PREFERENCE: Factor w/ 2 levels "For","Against": 1 1 1 1 1 1 2 1 1 1 ...

* 1. **Set the seed to 71923**

Ans **-** set.seed(71923)

* 1. 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.

Ans - The df\_train is the training data set with 70% of data and df\_test is the test data set with remaining 30% of the data

splitrule<-sample(nrow(dfk),0.7\*nrow(dfk))  
df\_train<-data.frame(dfk[splitrule,])  
df\_test<-data.frame(dfk[-splitrule,])

str(df\_train)

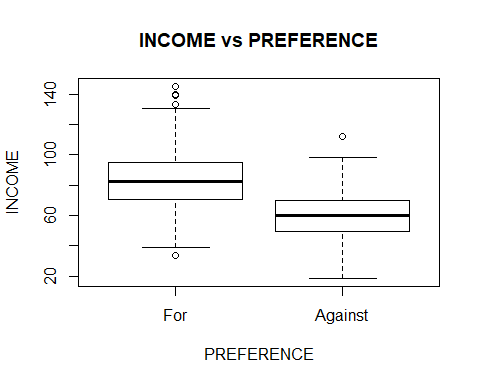
## 'data.frame': 700 obs. of 4 variables:  
## $ AGE : int 40 29 21 35 25 24 30 27 35 21 ...  
## $ INCOME : num 68 82.2 42.7 88.8 93.4 ...  
## $ GENDER : Factor w/ 2 levels "F","M": 2 2 2 1 2 2 2 1 2 2 ...  
## $ PREFERENCE: Factor w/ 2 levels "For","Against": 2 1 2 1 1 1 1 1 2 1 ...

str(df\_test)

## 'data.frame': 300 obs. of 4 variables:  
## $ AGE : int 50 26 27 39 23 15 25 31 40 45 ...  
## $ INCOME : num 113.2 107.5 80.9 83.8 53.2 ...  
## $ GENDER : Factor w/ 2 levels "F","M": 1 2 2 1 1 2 1 2 2 2 ...  
## $ PREFERENCE: Factor w/ 2 levels "For","Against": 1 1 1 1 1 1 2 1 2 1 ...

1. Exploratory analysis of the *training* data set
   1. Construct boxplots of INCOME and AGE (broken up by values of PREFERENCE). Present the plot as **Exhibit A**. What do you observe?

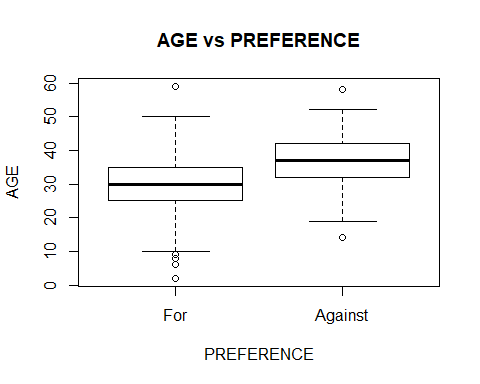
Ans -



The people who have the Preference “For” for the proposal have higher income as compared to those who are “Against” the proposal.

As per the above boxplot,

* The 3rd quartile of “For” is almost equal to the MAX of “Against”.
* The income of the “Against” outlier is lesser than the MAX of “For” income.
* The 3rd Quartile of “Against” is almost equal to the 1st Quartile of “For”.
* The number of outliers for “For” are more than that of “Against”.
* The median of “For” is higher than that of “Against”.



**Exhibit A**

The people who have the Preference “Against” for the proposal have higher age as compared to those who have the preference “For” the proposal.

As per the above boxplot,

* The minimum age for “For” is less than that of “Against”.
* Median of “For” is almost equal to the 1st quartile of “Against”.
* The number of outliers in “For” is more than that in “Against”.
  1. Construct a table for PREFERENCE showing proportions for and against.

Ans -

summ1<-summary(df\_train$PREFERENCE)  
prop.table(summ1)

## For Against   
## 0.8128571 0.1871429

Proportion is: For = 81.29%, Against = 18.71%

* 1. 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).

Ans-

num<-table(df\_train$GENDER,df\_train$PREFERENCE)  
num

##   
## For Against  
## F 276 76  
## M 293 55

Number of Women for the proposition = 276

Number of Women against the proposition = 76

Number of Men for the proposition = 293

Number of Men against the proposition = 55

1. Run a linear regression model of PREFERENCE on the demographic variables. Use only the training data set for fitting the model.

Ans –

df\_train$PREFERENCE<-revalue(df\_train$PREFERENCE,c("Against"="1","For"="0"))  
df\_train$PREFERENCE<-as.numeric(as.character(df\_train$PREFERENCE))  
fit1<-lm(df\_train$PREFERENCE~.,data=df\_train)  
summary(fit1)

##   
## Call:  
## lm(formula = df\_train$PREFERENCE ~ ., data = df\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.74013 -0.20850 -0.06941 0.16001 0.89611   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.3490527 0.0632336 5.520 4.79e-08 \*\*\*  
## AGE 0.0202591 0.0014690 13.791 < 2e-16 \*\*\*  
## INCOME -0.0096474 0.0005916 -16.308 < 2e-16 \*\*\*  
## GENDERM -0.0721760 0.0234616 -3.076 0.00218 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3102 on 696 degrees of freedom  
## Multiple R-squared: 0.371, Adjusted R-squared: 0.3683   
## F-statistic: 136.8 on 3 and 696 DF, p-value: < 2.2e-16

df\_test$PREFERENCE<-revalue(df\_test$PREFERENCE,c("Against"="1","For"="0"))  
df\_test$PREFERENCE<-as.numeric(as.character(df\_test$PREFERENCE))  
train.model<-lm(df\_train$PREFERENCE~.,df\_train)  
train.predict<-predict(train.model,df\_train)

test.predict<-predict(train.model,df\_test)

* 1. Compute the average error, RMSE and the mean absolute error (MAE) 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 for this yet).

Ans – For Training data set:

Average error = -9.061468e-16

RMSE = 0.3093274

MAE = 0.2429269

For Test data set:

Average error = 0.02349681

RMSE = 0.3239168

MAE = 0.261221

library(Metrics)

rmse(df\_train$PREFERENCE, train.predict)

## [1] 0.3093274

mean(df\_train$PREFERENCE - train.predict)

## [1] -9.061468e-16

mae(df\_train$PREFERENCE, train.predict)

## [1] 0.2429269

rmse(df\_test$PREFERENCE,test.predict)

## [1] 0.3239168

mean(df\_test$PREFERENCE - test.predict)

## [1] 0.02349681

mae(df\_test$PREFERENCE,test.predict)

## [1] 0.261221

* 1. For which data set are these errors smaller?

Ans – These errors are smaller for the training data set.

* 1. Use a cutoff of 0.5 and do the classification (i.e. make the class predictions). What proportions of predicted classes are for and against in each data set?

Ans – For training data set:

For = 0 is 90.29% of overall training data set

Against =1 is 9.71% of overall training data set

For Test data set:

For = 0 is 90.33% of overall training data set

Against = 1 is 9.67% of overall training data set

fit2<-lm(df\_train$PREFERENCE~.,data=df\_train)  
pred0<-ifelse(fit2$fitted.values>0.5,1,0)  
prop.table(table(pred0))

## pred0  
## 0 1   
## 0.90285714 0.09714286

predicted<-predict(fit2,newdata = df\_test)  
pred1<-ifelse(predicted>0.5,1,0)  
prop.table(table(pred1))

## pred1  
## 0 1   
## 0.90333333 0.09666667

* 1. What proportion of class predictions are in error in each of the training and test data set?

Ans – The proportion of class prediction that are in error for training data set is 1.71% + 10.71% = 12.42%.

The proportion of class prediction that are in error for test data set is 1.33% + 11.67% = 13%.

confusion0 <- table(df\_train$PREFERENCE, pred0)  
rownames(confusion0)<-c("For","Against")  
colnames(confusion0)<-c("For","Against")  
prop.table(confusion0)

## pred0  
## For Against  
## For 0.79571429 0.01714286  
## Against 0.10714286 0.08000000

confusion1<-table(df\_test$PREFERENCE,pred1)  
rownames(confusion1)<-c("For","Against")  
colnames(confusion1)<-c("For","Against")  
prop.table(confusion1)

## pred1  
## For Against  
## For 0.78666667 0.01333333  
## Against 0.11666667 0.08333333

1. Run a logistic regression model of PREFERENCE on the demographic variables. Use only the training data set for this.
   1. Present the output as **Exhibit B**.

Ans –

logfit<-glm(df\_train$PREFERENCE~.,data=df\_train,family="binomial")  
summary(logfit)

##   
## Call:  
## glm(formula = df\_train$PREFERENCE ~ ., family = "binomial", data = df\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.65726 -0.37215 -0.16886 -0.04293 2.76761   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.41788 0.76950 -0.543 0.58710   
## AGE 0.22478 0.02331 9.642 < 2e-16 \*\*\*  
## INCOME -0.11617 0.01129 -10.286 < 2e-16 \*\*\*  
## GENDERM -0.73941 0.27662 -2.673 0.00752 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 674.87 on 699 degrees of freedom  
## Residual deviance: 353.26 on 696 degrees of freedom  
## AIC: 361.26  
##   
## Number of Fisher Scoring iterations: 7

**Exhibit B**

* 1. Provide a precise interpretation of the coefficient of AGE.

Ans – 1 year increase in age increases the odds of having the Preference as “Against” by a factor of exp(0.224) = 1.25, for those with same Gender and income.

* 1. Provide a precise interpretation of the coefficient of the gender variable.

Ans – The coefficient of Gender = -0.739 i.e. exp(-0.739) = 0.477 implies that a increase in number of females decreases the odds of having the Preference as “Against” by a factor of 0.477for those with same age and income.

* 1. Use a cutoff of 0.5 and do the classification. What proportion of predicted classes are in error (in the training and test data set)?

Ans –

For training data set the proportion of predicted classes in error is 3.43% + 8% = 11.43%

For test data set the proportion of predicted classes in error is 3.33% + 8.33% = 11.66%

predictedTrain<-predict(logfit,newdata=df\_train,type="response")  
pred2<-ifelse(logfit$fitted.values>0.5,1,0)  
confusion2<-table(df\_train$PREFERENCE,pred2)  
rownames(confusion2)<-c("For","Against")  
colnames(confusion2)<-c("For","Against")

prop.table(confusion2)

## pred2  
## For Against  
## For 0.77857143 0.03428571  
## Against 0.08000000 0.10714286

predictedTest<-predict(logfit,newdata = df\_test,type="response")  
pred3<-ifelse(predictedTest>0.5,1,0)  
confusion3<-table(df\_test$PREFERENCE,pred3)  
rownames(confusion3)<-c("For","Against")  
colnames(confusion3)<-c("For","Against")  
prop.table(confusion3)

## pred3  
## For Against  
## For 0.76666667 0.03333333  
## Against 0.08333333 0.11666667

* 1. Compare these error rates with those in question 3d (linear regression).

Ans – The error rates in logistic regression are lesser than those in linear regression.

* 1. 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.

Ans – The predicted probability for voting against the proposition given the above values is 0.3875589

dftester<-data.frame(AGE= 36,INCOME= 70,GENDER= "F")  
dftester$prob<-predict(logfit,newdata = dftester,type="response")  
dftester[,c(4)]

## [1] 0.3875589

**Hints:** You may find the *R***ifelse** function convenient for classification. Finally, the **predict** function that was used for regression will also work for the logistic case. Note however that, by default, 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")**