Homework 6

Prakash Paudyal

December 10, 2017

Please do the following problems from the text book ISLR.

1.** $\boldsymbol{Question 5.4.3 pg 198}$ **

3. We now review k-fold cross-validation. (a) Explain how k-fold cross-validation is implemented.

Ans: In the k-fold cross-validation, the data is randomly divided into K equal parts (k-folds). One of its part is used as validation data for testing the model and other k-1 parts of data are used as training data to fit the model. Then the MSE is calculated for held out data. Then this process is repeated for k times, each time, different parts of data is used as validation set. The test error is the estimated by averaging k numbers of MSE's. In this method all the data observations are used as both training and test set.

$$CV_(k) = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

(b) What are the advantages and disadvantages of k-fold crossvalidation relative to:

i. The validation set approach?

Advantages: K-fold CV test error rate can be less variable than the validation set approach, since training set contains more observations. K-fold CV has more accuracy on estimates of the test error.

Disadvantage: K-fold CV need more computatioanl cost and is complex to implement than validation set approach

ii. LOOCV?

Advantages:

If n is large, k-folds cross validation with k less than n provides a much more cost-effective and computationally efficient estimate. K-fold cross validation with k less than n often gives more accurate estimates of the test error rate than does LOOCV. K-fold cross validation with k less than n has also lower vriance than LOOCV.

Disadvantage: K-fold cross validation with k less than n may give biased estimates of test error compared to the LOOCV cross-validation approach which gives approximately unbiased estimates of the test error, since each training set contains n-1 observations.

2. *Question* 5. 4. 5*pg* 198(*useset*. *seed*(702)*tomakeresultsreplicable*)

5. In Chapter 4, we used logistic regression to predict the probability of default using income and balance on the Default data set. We will now estimate the test error of this logistic regression model using the validation set approach. Do not forget to set a random seed before beginning your analysis.

(a) Fit a logistic regression model that uses income and balance to predict default.

```
## Warning: package 'ISLR' was built under R version 3.4.1
```

```
##
## Call:
## glm(formula = default ~ income + balance, family = "binomial",
##     data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4725  -0.1444  -0.0574  -0.0211   3.7245
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.154e+01  4.348e-01 -26.545  < 2e-16 ***
## income       2.081e-05  4.985e-06   4.174 2.99e-05 ***
## balance      5.647e-03  2.274e-04  24.836  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1579.0  on 9997  degrees of freedom
## AIC: 1585
##
## Number of Fisher Scoring iterations: 8
```

(b) Using the validation set approach, estimate the test error of this model. In order to do this, you must perform the following steps:

i. Split the sample set into a training set and a validation set.

```
## 50% of the sample size
smp_size <- floor(0.50 * nrow(Default))
## set the seed to make partition reproductible
set.seed(702)
#train_ind1<- sample(dim(Default)[1], dim(Default)[1] / 2)
train_ind <- sample(seq_len(nrow(Default)), size = smp_size)

train <- Default[train_ind, ]
test <- Default[-train_ind, ]
```

ii. Fit a multiple logistic regression model using only the training observations.

```
##
## Call:
## glm(formula = default ~ income + balance, family = "binomial",
##     data = Default, subset = train_ind)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5177  -0.1316  -0.0484  -0.0170   3.6037
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.192e+01  6.589e-01 -18.085   <2e-16 ***
## income       1.465e-05  7.191e-06   2.037   0.0417 *
## balance      5.983e-03  3.566e-04  16.780   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1375.12  on 4999  degrees of freedom
## Residual deviance:  730.94  on 4997  degrees of freedom
## AIC: 736.94
##
## Number of Fisher Scoring iterations: 8
```

iii. Obtain a prediction of default status for each individual in the validation set by computing the posterior probability of default for that individual, and classifying the individual to the default category if the posterior probability is greater than 0.5.

```
## iii
probs <- predict(train.glm, newdata = test, type = "response")
pred.glm <- rep("No", length(probs))
pred.glm[probs > 0.5] <- "Yes"
```

iv. Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.

```
## [1] 0.0274
```

Disucssion: 2.74% of validation sets were misclassified.

(c) Repeat the process in (b) three times, using three different splits of the observations into a training set and a validation set. Comment on the results obtained.

60% of the sample size

```
##
## Call:
## glm(formula = default ~ income + balance, family = "binomial",
##     data = Default, subset = train_ind)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4579  -0.1425  -0.0571  -0.0211   3.7393
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.125e+01  5.426e-01 -20.735   <2e-16 ***
## income       1.173e-05  6.372e-06   1.841   0.0657 .
## balance      5.701e-03  2.872e-04  19.851   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1847.03  on 5999  degrees of freedom
## Residual deviance:  961.07  on 5997  degrees of freedom
## AIC: 967.07
##
## Number of Fisher Scoring iterations: 8
```

70% of the sample size

```
##
## Call:
## glm(formula = default ~ income + balance, family = "binomial",
##     data = Default, subset = train_ind)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4018  -0.1526  -0.0611  -0.0235   3.6745
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.107e+01  4.948e-01  -22.37  < 2e-16 ***
## income       1.625e-05  5.868e-06    2.77  0.00561 **
## balance      5.469e-03  2.622e-04   20.86  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2064.0  on 6999  degrees of freedom
## Residual deviance: 1145.6  on 6997  degrees of freedom
## AIC: 1151.6
##
## Number of Fisher Scoring iterations: 8
```

80% of the sample size

```
##
## Call:
```

```
## glm(formula = default ~ income + balance, family = "binomial",
##     data = Default, subset = train_ind)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4639  -0.1442  -0.0577  -0.0210   3.7277
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.155e+01  4.830e-01 -23.918  < 2e-16 ***
## income       2.101e-05  5.561e-06   3.778 0.000158 ***
## balance      5.639e-03  2.521e-04  22.368  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2354.0  on 7999  degrees of freedom
## Residual deviance: 1262.7  on 7997  degrees of freedom
## AIC: 1268.7
##
## Number of Fisher Scoring iterations: 8
```

Models	sampleSize	ErrorRate
Model_1	60%	0.0253
Model_2	70%	0.0250
Model_3	80%	0.0275

***Discussion**:* I used 60%,70% and 80% sample size of observations of data to train models. The validation estimate of the test error rates are varied, depending on precisely which observations are included in the training set and which observations are included in the validation set. Splitting data into 80 :20 ratio gave best result.

(d) Now consider a logistic regression model that predicts the probability of default using income, balance, and a dummy variable for student. Estimate the test error for this model using the validation set approach. Comment on whether or not including a dummy variable for student leads to a reduction in the test error rate.

```
##
## Call:
## glm(formula = default ~ income + balance + student, family = "binomial",
##     data = Default, subset = train_ind)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4617  -0.1441  -0.0572  -0.0209   3.7243
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -1.128e+01  5.582e-01 -20.202   <2e-16 ***
## income       1.396e-05  9.203e-06   1.517    0.129
## balance      5.669e-03  2.549e-04  22.243   <2e-16 ***
## studentYes  -2.555e-01  2.655e-01  -0.962    0.336
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2354.0  on 7999  degrees of freedom
## Residual deviance: 1261.8  on 7996  degrees of freedom
## AIC: 1269.8
##
## Number of Fisher Scoring iterations: 8

## [1] 0.0275
```

Discussion I used 80% sample of data to train model as in Q2.c.Both models produces same validation test error=0.0275. Hence, we can say that adding student as predictor in model does't help model to reduce the test error.

3. *Question* 5.4.7 *pg* 200

7. In Sections 5.3.2 and 5.3.3, we saw that the cv.glm() function can be used in order to compute the LOOCV test error estimate. Alternatively,one could compute those quantities using just the glm() and predict.glm() functions, and a for loop. You will now take this approach in order to compute the LOOCV error for a simple logistic regression model on the Weekly data set. Recall that in the context of classification problems, the LOOCV error is given in (5.4).

(a) Fit a logistic regressionmodel that predicts Direction using Lag1 and Lag2.

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2, family = "binomial", data = Weekly)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.623  -1.261   1.001   1.083   1.506
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.22122    0.06147   3.599 0.000319 ***
## Lag1        -0.03872    0.02622  -1.477 0.139672
## Lag2         0.06025    0.02655   2.270 0.023232 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1488.2  on 1086  degrees of freedom
## AIC: 1494.2
##
## Number of Fisher Scoring iterations: 4
```

(b) Fit a logistic regressionmodel that predicts Direction using Lag1 and Lag2 using all but the first observation.

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2, family = "binomial", data = Weekly[
-1,
##     ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6258  -1.2617   0.9999   1.0819   1.5071
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.22324    0.06150   3.630 0.000283 ***
## Lag1        -0.03843    0.02622  -1.466 0.142683
## Lag2         0.06085    0.02656   2.291 0.021971 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1494.6  on 1087  degrees of freedom
## Residual deviance: 1486.5  on 1085  degrees of freedom
## AIC: 1492.5
##
## Number of Fisher Scoring iterations: 4
```

(c) Use the model from (b) to predict the direction of the first observation. You can do this by predicting that the first observation will go up if P(Direction="Up"|Lag1, Lag2) > 0.5. Was this observation correctly classified?

```
##    1
## TRUE
```

Yes observation was correctly classified.

(d) Write a for loop from i = 1 to i = n, where n is the number of observations in the data set, that performs each of the following steps:

i. Fit a logistic regression model using all but the ith observation to predict Direction using Lag1 and Lag2.

ii. Compute the posterior probability of the market moving up for the ith observation.

iii. Use the posterior probability for the ith observation in order to predict whether or not the market moves up.

iv. Determine whether or not an error was made in predicting the direction for the ith observation. If an error was made,then indicate this as a 1, and otherwise indicate it as a 0.

```
error <- rep(0, dim(Weekly)[1])
for (i in 1:dim(Weekly)[1]) {
  fit.glm <- glm(Direction ~ Lag1 + Lag2, data = Weekly[-i, ],  family = "bin
omial")
  pred.up <- predict.glm(fit.glm, Weekly[i, ], type = "response") > 0.5
  true.up <- Weekly[i, ]$Direction == "Up"
  if (pred.up != true.up)
    error[i] <- 1
}
error
```

```
##    [1] 1 1 0 1 0 1 0 0 0 1 1 0 1 0 1 0 1 0 1 0 0 0 1 1 1 1 1 1 0 1 1 1 1 0
##   [35] 1 0 0 0 1 0 1 0 0 1 0 1 1 1 0 1 0 0 0 1 0 0 1 1 0 0 0 0 1 0 1 1 0 0
##   [69] 1 0 1 1 0 0 0 1 0 1 1 0 0 1 1 0 1 1 0 0 1 0 0 1 1 1 0 0 0 0 0 1 0 1
##  [103] 1 0 0 1 0 1 0 0 1 1 0 0 1 0 0 1 0 0 1 1 1 1 0 0 0 1 0 1 0 1 1 0 0 0
##  [137] 1 1 1 0 0 0 1 0 0 0 0 0 0 1 1 1 0 1 0 0 1 1 0 1 0 0 1 1 0 0 1 0 0 1
##  [171] 0 0 1 1 1 0 1 0 1 0 0 0 0 0 0 0 0 1 1 0 1 0 1 0 1 0 1 0 0 1 0 0 1 0
##  [205] 0 1 0 1 0 1 1 1 0 0 1 1 0 1 0 0 1 1 0 0 0 1 1 1 0 1 0 1 0 1 0 0 0 1
##  [239] 1 0 1 0 1 0 1 0 1 0 1 1 0 1 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 0 1 0 0 1 0
##  [273] 0 0 1 0 0 1 0 0 1 0 0 1 0 1 1 0 0 0 0 0 1 0 1 0 0 1 0 0 0 1 0 0 1 1
##  [307] 0 0 1 0 0 0 0 1 0 1 1 0 0 1 0 1 0 1 1 0 0 0 1 0 1 0 0 1 1 1 1 0 1 0
##  [341] 0 1 0 0 0 1 0 1 0 1 0 0 0 0 0 1 1 0 0 1 0 0 1 0 0 0 1 1 0 1 1 1 1 1
##  [375] 0 0 0 1 0 0 0 0 0 0 1 0 1 1 0 0 1 1 0 0 0 0 0 1 0 0 1 1 1 0 1 0 1 0
##  [409] 1 1 1 0 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 1 0 0 1 0 1 0 0 0 0 0 1 1
##  [443] 1 1 0 1 1 0 1 0 1 1 0 1 0 0 1 0 0 1 1 0 0 0 0 1 1 0 0 1 0 1 0 0 0 1
##  [477] 0 0 1 0 0 0 1 1 1 0 1 0 0 0 1 0 1 1 1 0 0 0 0 1 1 1 0 1 1 0 1 0 0 0
##  [511] 1 0 1 0 0 0 1 0 1 1 0 0 1 1 0 0 0 1 1 0 1 0 1 1 1 1 1 0 0 0 1 0 0 0
##  [545] 1 1 0 1 0 0 0 1 1 1 1 1 1 0 1 0 1 0 0 1 0 0 1 1 1 0 0 0 1 1 1 1 1 1
##  [579] 1 1 1 1 0 1 0 0 1 0 0 1 0 1 0 0 1 0 0 1 0 1 1 0 1 1 1 0 1 0 1 0 1 0
##  [613] 0 0 1 0 1 0 1 0 1 1 0 1 1 0 1 0 1 0 1 1 1 1 0 1 1 0 0 0 1 1 1 1 0 1
##  [647] 1 1 0 1 0 0 0 1 1 1 1 1 1 0 1 0 0 1 0 0 0 1 1 0 1 0 1 1 1 1 0 0 0 1
##  [681] 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 1 0 1 1 0 0 0 0 1 0 1 0 1 0 1
##  [715] 0 0 1 1 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 1 1 1 0 1 1 0 1 1 1 1 0 0 0 1
##  [749] 1 1 1 1 1 0 1 0 0 0 0 0 0 1 0 1 1 1 0 0 0 1 0 0 1 0 0 0 1 1 1 0 0 0
##  [783] 1 0 0 1 1 1 0 0 1 0 1 0 1 0 0 1 0 0 1 0 0 0 0 0 1 0 1 1 0 0 1 1 0 1
##  [817] 1 1 0 0 0 0 0 1 1 0 0 1 0 0 1 0 1 0 0 0 1 1 0 1 1 0 1 0 1 0 1 1 0 0
##  [851] 1 1 1 0 1 1 0 0 0 1 0 1 0 1 0 1 0 0 0 0 0 1 0 0 1 1 0 0 1 0 1 0 1 1
##  [885] 0 1 0 1 1 0 0 0 1 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 1 1 1 1 1 0 1 1 0 0
##  [919] 0 0 0 1 0 0 1 0 0 0 0 1 0 1 1 1 0 0 1 1 0 1 1 1 1 0 1 0 1 0 1 0 1 0
##  [953] 1 0 0 1 1 1 1 1 1 0 1 0 0 0 1 1 1 0 1 1 1 1 0 0 0 0 1 1 0 0 0 0 1 0 0
##  [987] 1 1 1 0 0 1 1 1 0 1 0 0 0 0 1 0 0 1 0 1 0 0 1 1 1 1 0 1 0 0 1 0 0 1
## [1021] 0 0 1 1 0 1 1 1 0 1 1 0 0 0 1 0 1 0 1 1 1 1 0 0 1 0 0 0 0 0 0 1 0 1
## [1055] 1 0 0 0 0 0 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0
## [1089] 0
```

(e) Take the average of the n numbers obtained in (d)iv in order to obtain the LOOCV estimate for the test error. Comment on the results.

```
## [1] 0.4499541
```

LOOCV estimates the 44.99% test error.

4. Write your own code (similar to Q #3. above) to estimate test error using 6-fold cross validation for fitting linear regression with from the Auto data in the ISLR library.

ANS:

Here I implemented this equation $CV_(k) = \frac{1}{k}\sum_{i=1}^{k} MSE_i$ for cross validation. I have split data into 6 equal parts with out replacement.

```
library(knitr)
library(plyr)
library(ISLR)
data(Auto)
#Auto$mpg01 <- ifelse(Auto$mpg>median(Auto$mpg),1,0)
attach(Auto)
set.seed(1272)
folds <- split(Auto, cut(sample(1:nrow(Auto),replace = FALSE),6))
errs <- rep(NA, length(folds))
for (i in 1:length(folds)) {
 test <- ldply(folds[i], data.frame)
 train <- ldply(folds[-i], data.frame)
 lm_train <- lm(mpg ~ horsepower + I(horsepower^2), data = train)
  errs [i]<- mean((test$mpg - predict(lm_train,test))^2)
  }
fold<-c("fold1","fold2","fold3","fold4","fold5","fold6")
MSES<-c("MSE1","MSE2","MSE3","MSE4","MSE5","MSE6")
d<-data.frame(fold,MSES, MSE=errs)
kable(d)
```

fold	MSES	MSE
fold1	MSE1	19.50502
fold2	MSE2	19.48914
fold3	MSE3	18.63261
fold4	MSE4	17.58160
fold5	MSE5	18.03205
fold6	MSE6	22.67756

```
mean(errs)
```

```
## [1] 19.31966
```

5. Last homework you started analyzing the dataset you chose from this website. Now continue the analysis and perform Logistic Regression, KNN, LDA, QDA, MclustDA, MclustDA with EDDA if appropriate. If it is not possible to perform any of the methods mentioned above please justify why.

Discussion I have chossen the adult income which has 32561 observations and 15 variables .I tried to predict the income of the person according to differet predictors liks age“,”workclass“,”fnlwgt“,”education“,”education-num“,”marital-status“,”occupation“,”relationship“,”race“,”sex“,”capitalgain“,”capitalloss“,”hoursperweek“,” nativecountry“,”income". Since data observation was big, I subseted data for the age greater than 30 to estimate either his or her income comes under 50k are above 50k. For this I chose some of the variables I thought they are good predictor for income. I fit the model for glm, qda and knn=1 but I could not fit qda and MclustDA. I could not figure it out this time why I was not able to fit some models. I found this problem was intresting to work on raw data,I will continue the work on this problem.If I found somethig , I will report you.

	age	education	occupation	hoursperweek	sex	income	income1
1	39	Bachelors	Adm-clerical	40	Male	<=50K	0
2	50	Bachelors	Exec-managerial	13	Male	<=50K	0
3	38	HS-grad	Handlers-cleaners	40	Male	<=50K	0
4	53	11th	Handlers-cleaners	40	Male	<=50K	0
6	37	Masters	Exec-managerial	40	Female	<=50K	0
7	49	9th	Other-service	16	Female	<=50K	0

5.GLM

```
##
## Call:
## glm(formula = income1 ~ age + education + occupation + hoursperweek,
##     family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4373  -0.8194  -0.5131   0.9477   2.7243
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -4.626639   0.230195 -20.099  < 2e-16 ***
## age                            0.017997   0.001830   9.835  < 2e-16 ***
## education 11th                -0.175938   0.235588  -0.747   0.4552
## education 12th                 0.501790   0.272044   1.845   0.0651 .
## education 1st-4th             -0.871029   0.489840  -1.778   0.0754 .
## education 5th-6th             -0.402826   0.340914  -1.182   0.2374
## education 7th-8th             -0.481798   0.249203  -1.933   0.0532 .
## education 9th                 -0.301582   0.279749  -1.078   0.2810
## education Assoc-acdm           1.117562   0.183895   6.077 1.22e-09 ***
```

```
## education Assoc-voc             1.281617   0.177009   7.240 4.47e-13 ***
## education Bachelors             1.827454   0.164546  11.106  < 2e-16 ***
## education Doctorate             2.755741   0.211978  13.000  < 2e-16 ***
## education HS-grad               0.741973   0.160865   4.612 3.98e-06 ***
## education Masters               2.057314   0.172687  11.914  < 2e-16 ***
## education Preschool           -10.900089  86.698892  -0.126   0.9000
## education Prof-school           3.037570   0.205734  14.765  < 2e-16 ***
## education Some-college          1.054592   0.163147   6.464 1.02e-10 ***
## occupation Adm-clerical         0.116819   0.119158   0.980   0.3269
## occupation Armed-Forces         1.117841   1.476318   0.757   0.4489
## occupation Craft-repair         0.794110   0.114059   6.962 3.35e-12 ***
## occupation Exec-managerial      1.250342   0.111393  11.225  < 2e-16 ***
## occupation Farming-fishing     -0.289520   0.160393  -1.805   0.0711 .
## occupation Handlers-cleaners   -0.059950   0.175427  -0.342   0.7325
## occupation Machine-op-inspct    0.325078   0.132718   2.449   0.0143 *
## occupation Other-service       -0.757080   0.148414  -5.101 3.38e-07 ***
## occupation Priv-house-serv     -2.245696   1.016156  -2.210   0.0271 *
## occupation Prof-specialty       0.757346   0.114038   6.641 3.11e-11 ***
## occupation Protective-serv      0.988551   0.153221   6.452 1.11e-10 ***
## occupation Sales                0.827780   0.115146   7.189 6.53e-13 ***
## occupation Tech-support         1.033704   0.140463   7.359 1.85e-13 ***
## occupation Transport-moving     0.534963   0.131190   4.078 4.55e-05 ***
## hoursperweek                    0.029873   0.001676  17.821  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22244  on 17590  degrees of freedom
## Residual deviance: 18414  on 17559  degrees of freedom
## AIC: 18478
##
## Number of Fisher Scoring iterations: 12

## [1] 0.2528422
```

5.knn

```
##
## knn.pred    0    1
##        0 2787  241
##        1  210 1160

## [1] 0.1025466
```

5.lda

```
##         Length Class  Mode
## prior    2     -none- numeric
```

```
## counts   2     -none- numeric
## means   62     -none- numeric
## scaling 31     -none- numeric
## lev      2     -none- character
## svd      1     -none- numeric
## N        1     -none- numeric
## call     3     -none- call
## terms    3     terms  call
## xlevels  2     -none- list

## [1] 0.7460209
```

5errorrate

glmerror	ldaerror	knnerror
0.2528422	0.7460209	0.1025466

KNN provides less test error compared to other models to predict the income class.