Assignment 6
Shaheed Shihan
Modern Applied Statistics

3. We now review k-fold cross validation.
   a) Explain how k-fold cross-validation is implemented.
   b) What are the advantages and disadvantages of k-fold cross validation relative to:
      i. The validation set approach?
      ii. LOOCV?

3. a) The k-fold cross validation is implemented by taking a certain number of observations, let's say n, then randomly splitting it into non-overlapping groups (k number of groups) of length n/k. This group is the validation set, and whatever is left (n-n/k) is the training set. The test error is found by averaging the k resulting MSE estimates.

i) Depending on which observations are included and which aren't, the test error estimate of the validation set can be highly variable. Also, the test error rate in the validation set approach tends to be overestimated as compared to the k-fold. This is because the statistical methods tend to produce bad results when implemented on a small training set.
ii) In comparison, the LOOCV approach is much better. There is usually far less bias because the in the LOOCV approach we repeatedly fit the statistical learning method using the training sets that contain n-1 observations, which is almost as large as the entire data set. The LOOCV approach tends not to overestimate the test error rate. However, variance of LOOCV can be higher than k-fold and LOOCV can be computationally very expensive and time consuming to process. There is a bias-variance trade off when choosing the k in the k-fold approach.

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

```
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).

```
Call:
glm(formula = default ~ income + balance, family = "binomial",
    data = d.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4221  -0.1448  -0.0571  -0.0211   3.7346

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.142e+01  5.410e-01 -21.118  < 2e-16 ***
income       1.601e-05  6.127e-06   2.612  0.00899 **
balance      5.645e-03  2.862e-04  19.726  < 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: 1865.5  on 6665  degrees of freedom
Residual deviance: 1030.8  on 6663  degrees of freedom
AIC: 1036.8

Number of Fisher Scoring iterations: 8

> mean(glm2pred!=d.test$default)
[1] 0.02909418
```

(c)

```
> mean(glm2pred!=d.test$default)
[1] 0.02729454
> mean(glm2pred!=d.test$default)
```

[1] 0.02669466

> mean(glm2pred!=d.test$default)
[1] 0.02609478

The means stayed the same for the most part.

(d)

```
Call:
glm(formula = default ~ income + balance + student, family = "binomial",
    data = d.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5400  -0.1418  -0.0545  -0.0196   3.7437

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.134e+01  6.251e-01 -18.145   <2e-16 ***
income       1.145e-05  1.021e-05   1.121    0.262
balance      5.835e-03  2.943e-04  19.825   <2e-16 ***
studentYes  -4.519e-01  2.945e-01  -1.534    0.125
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1913.3  on 6665  degrees of freedom
Residual deviance: 1035.9  on 6662  degrees of freedom
AIC: 1043.9

Number of Fisher Scoring iterations: 8

> mean(glm2pred!=d.test$default)
[1] 0.02789442
```

The test error rate stayed the same for the most part even when including a dummy variable for student.

7.
(a)

```
Call:
```

```
glm(formula = Direction ~ Lag1 + Lag2, family = "binomial", data = W)

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)

```
Call:
glm(formula = Direction ~ Lag1 + Lag2, family = "binomial", data = W[-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)

```
> predict.glm(glm.W2, Weekly[1, ], type = "response") > 0.5
   1 
TRUE 

> head(W)
  Year   Lag1   Lag2   Lag3   Lag4   Lag5    Volume  Today Direction
1 1990  0.816  1.572 -3.936 -0.229 -3.484 0.1549760 -0.270      Down
2 1990 -0.270  0.816  1.572 -3.936 -0.229 0.1485740 -2.576      Down
```

Therefore, our model misclassified the direction wrong.

(d)

```
> 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 0 1 1 1 1 0 1 0 0 0 1 0 1 0 0
  [44] 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 1 0 1 1 0 0 0 1 0 1 1 0 0 1 1 0 1 1
  [87] 0 0 1 0 0 1 1 1 0 0 0 0 0 1 0 1 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
 [130] 1 0 1 1 0 0 0 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 0 0
 [173] 1 1 1 0 1 0 1 0 0 0 0 0 0 0 1 1 0 1 0 1 0 1 0 1 0 0 1 0 0 1 0 0 1 0 1 0 1 1 1 0 0 1
 [216] 1 0 1 0 0 1 1 0 0 0 1 1 1 0 1 0 1 0 1 0 0 0 1 1 0 1 0 1 0 1 0 1 0 1 1 0 1 0 0 1 0 0 1
 [259] 0 0 0 0 0 1 0 0 0 1 0 0 1 0 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
 [302] 1 0 0 1 1 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 0 1 0 0
 [345] 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 0 0 0 1 0 0 0 0 0 0 1 0 1
 [388] 1 0 0 1 1 0 0 0 0 0 1 0 0 1 1 1 0 1 0 1 0 1 1 1 0 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 1
 [431] 0 0 1 0 1 0 0 0 0 0 1 1 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
 [474] 0 0 1 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 1 0 1 0 0 0
 [517] 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 1 1 0 1 0 0 0 1 1 1 1 1 1 0 1
 [560] 0 1 0 0 1 0 0 1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 0 1 0 0 1 0 0 1 0 1 0 0 1 0 0 1 0 1 1 0
 [603] 1 1 1 0 1 0 1 0 1 0 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
 [646] 1 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 1 0 0 0 0 0 0 0
 [689] 0 0 0 1 0 0 0 0 1 0 0 1 0 1 1 0 0 0 0 1 0 1 0 1 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 1 0 0 1
 [732] 0 0 1 1 1 0 1 1 0 1 1 1 1 0 0 0 1 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
 [775] 0 0 1 1 1 0 0 0 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 1
 [818] 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 1 1 1 0 1 1 0 0 0 1
 [861] 0 1 0 1 0 1 0 0 0 0 0 1 0 0 1 1 0 0 1 0 1 0 1 1 0 1 0 1 1 0 0 0 1 0 0 0 0 0 0 0 1 0 1
 [904] 0 1 0 0 0 1 1 1 1 1 0 1 1 0 0 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
 [947] 1 0 1 0 1 0 1 0 0 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 1 1 1
 [990] 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 0 0 1 1 0 1 1 1 0 1 1 0
[1033] 0 0 1 0 1 0 1 1 1 1 0 0 1 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0
```

```
[1076] 1 0 0 0 0 0 1 0 1 0 0 0 0 0
```

(e)
> mean(error)
[1] 0.4499541

4.
> cv.error.6
[1] 0.1395584 0.1215767 0.1211042 0.1179214 0.1152967 0.1168857

The average error with a 6-fold cross validation was around 12.21%.

5.
I kept getting an error for my models. I will run it again and figure it out.