Practice Exam #4

No notes, calculators, or R programming will be allowed during this exam. No use of R is required to complete the questions below. The exam will be short enough for students to complete within the allotted 2 hours.

Case Study #1

You are the CHRO of Kramerica Industries, a consulting firm. You are tasked with increasing employee productivity AND improving hiring practices over the next eighteen months. Use the dataset described below to answer the questions and develop a plan of action for each. The appendix has all of the information you'll need to answer each question.

Variable	Description
technical	1 indicates this employee has a technical background, 0 otherwise (0 could be an HR role, an administrative role, etc.)
yearsofservice	number of years the employee has worked for the firm (0 indicates a new, entry-level employee)
currentsalary	total annual salary for each employee at the firm
performancereview	values of 1-10 with 10 being an excellent review at the end of last year
leadershiplevel	values of 1-5 where 5 is the highest level of promotion and 1 is entry level
levelofeducation	values of 1-5 where 5 is PhD or similar, 4 is MS, MBA or similar, 3 is college graduate, 2 is some college, and 1 is high school graduate
certifications	number of professional certifications held by employee
peerreviews	values of 1-10 with 10 being an excellent peer review at the end of last year

1. In testing the performance of this model, how should the data be divided into training/test sets?
2. Do we need to worry about outliers for this model?
3. What do we look for when comparing the errors in the training set to the errors in the test set?
4. What should we do if the errors are much larger on average in the test set than in the training set?

Case Study #2

You are the Operations Manager of FedEx distribution centers in the US. In an effort to improve daily delivery efficiency, you've asked the Operations Analytics team to create a couple of models for you. The models are included in the appendices. The data used is described below.

Variable	Description
driversworking	total number of drivers employed by this firm who are delivering packages on this date
weekend	1 indicates this observation is on a weekend, 0 otherwise
expectedpackagesdelivered	total number of packages planned for delivery on this date
extrahands	1 if an additional 1,000 workers should have been hired

	temporarily for this day
weatherconditions	100% indicates perfect weather, 0% indicates bad weather (snow, no packages delivered)
pctoversized	percent of packages that are oversized on this date

5. Why do we sometimes include interaction terms in a model?
6. Why do we sometimes include nonlinear terms in a model?
7. Interpret the interaction terms in Appendix 2, if any.
8. Interpret the nonlinear terms in Appendix 2, if any.
9. What type of model should we create to predict how many drivers should be working on a given day?
10. What type of model should we create to predict whether or not we need extra hands on a given day?
11. Based on Appendix 3, does the model predict as well out of sample as it does in sample? (Is the model stable?
12. Based on Appendix 3, and specifically the confusion matrix of the test set, how often is this model correct in its predictions?
13. Based on Appendix 3, and specifically the confusion matrix of the test set, how often is the model incorrect in its predictions?
14. Based on Appendix 3, and specifically the confusion matrix of the test set, what could be the economic impact when the model incorrectly predicts 0?
15. Based on Appendix 3, and specifically the confusion matrix of the test set, what could be the economic impact when the model incorrectly predicts 1?

There is no appendix to help answer these questions, but these may appear on the exam:

16. What can a decision tree do?
17. How many types of statistical decision trees are there?
18. Compare two error distributions and choose whether you would prefer to use a decision tree or a linear regression for this problem.
19. Which model should you choose if you want to understand relationships between predictors and a continuous response? Any words of caution? (Hint: First decide which models you have to choose from.)
20. Which model should you choose if you want to predict outcomes of a continuous response? Any words of caution? (Hint: First decide which models you have to choose from.)
21. Compare two confusion matrices and choose whether you would prefer to use a decision tree or a logistic regression for this problem based on their results.
22. Which model should you choose if you want to understand relationships between predictors and a binary response? Any words of caution? (Hint: First decide which models you have to choose from.)
23. Which model should you use if you want to predict outcomes of a binary response? Any words of caution? (Hint: First decide which models you have to choose from.)
24. What issues might I run into when using a decision tree model that I don't run into when I use a linear regression or logistic regression model?
25. What issues might I run into when using a linear or logistic regression model that I don't run into when I use a decision tree?

Appendix 1

```
Call:
lm(formula = d$performancereview ~ d$yearsofservice + d$currentsalary +
    d$levelofeducation + d$certifications + d$peerreviews)

Residuals:
     Min       1Q   Median       3Q      Max
-2.51603 -0.51354 -0.02218  0.51669  2.75217

Coefficients:
                      Estimate    Std. Error t value             Pr(>|t|)
(Intercept)       -0.7835507060  0.2064429424  -3.795             0.000155 ***
d$yearsofservice   0.0268004653  0.0139700211   1.918             0.055295 .
d$currentsalary    0.0000070797  0.0000009722   7.282  0.00000000000059598 ***
d$levelofeducation 0.1760585259  0.0222372075   7.917  0.00000000000000551 ***
d$certifications  -0.1058992194  0.0309269703  -3.424             0.000638 ***
d$peerreviews      0.6536765507  0.0258367315  25.300 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7877 on 1194 degrees of freedom
Multiple R-squared:  0.5143,  Adjusted R-squared:  0.5123
F-statistic: 252.9 on 5 and 1194 DF,  p-value: < 0.00000000000000022
```

```
> cor(cbind(perf=d$performancereview,tech=d$technical,years=d$yearsofservice,salary=d
$currentsalary,edu=d$levelofeducation,cert=d$certifications,peer=d$peerreviews))
              perf          tech       years    salary         edu       cert          peer
perf    1.0000000000 -0.0002313757 0.356684895 0.4724416  0.12752214 0.03138277  0.6499061630
tech   -0.0002313757  1.0000000000 0.000508196 0.7456322  0.02701159 0.23644815  0.0001447307
years   0.3566848951  0.0005081960 1.000000000 0.5195143  0.16716914 0.32432113  0.3438252505
salary  0.4724415765  0.7456322037 0.519514350 1.0000000  0.28575744 0.23509944  0.3786013113
edu     0.1275221367  0.0270115906 0.167169141 0.2857574  1.00000000 0.07796352 -0.1761959757
cert    0.0313827656  0.2364481469 0.324321132 0.2350994  0.07796352 1.00000000  0.0525639547
peer    0.6499061630  0.0001447307 0.343825251 0.3786013 -0.17619598 0.05256395  1.0000000000
```

Appendix 2

```
> weekendpackagesint = dtrn$weekend*dtrn$expectedpackagesdelivered
> fit = lm(dtrn$driversworking ~ dtrn$weekend + dtrn$expectedpackagesdelivered + weekendpackagesint)
> summary(fit)

Call:
lm(formula = dtrn$driversworking ~ dtrn$weekend + dtrn$expectedpackagesdelivered +
    weekendpackagesint)

Residuals:
     Min       1Q   Median       3Q      Max
-20929.8  -4211.9   -285.5   4339.0  17832.8

Coefficients:
                                 Estimate     Std. Error t value             Pr(>|t|)
(Intercept)                  111111.7145399   4294.7165259  25.872 < 0.0000000000000002 ***
dtrn$weekend                  26845.4503646   9010.9893219   2.979              0.00303 **
dtrn$expectedpackagesdelivered      0.0077472      0.0003654  21.200 < 0.0000000000000002 ***
weekendpackagesint                 -0.0025925      0.0008218  -3.155              0.00170 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6489 on 506 degrees of freedom
Multiple R-squared:  0.5729,  Adjusted R-squared:  0.5704
F-statistic: 226.2 on 3 and 506 DF,  p-value: < 0.0000000000000022
```

Appendix 3

```
> fit = glm(dtrn$extrahands ~ dtrn$expectedpackagesdelivered + dtrn$driversworking + dtrn
$weatherconditions + dtrn$pctoversized,family="binomial")
> summary(fit)

Call:
glm(formula = dtrn$extrahands ~ dtrn$expectedpackagesdelivered +
    dtrn$driversworking + dtrn$weatherconditions + dtrn$pctoversized,
    family = "binomial")

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.34292  -0.38748  -0.17271  -0.05092   2.57307

Coefficients:
                                  Estimate     Std. Error z value             Pr(>|z|)
(Intercept)                  -14.4992196310   3.8479423169  -3.768             0.000165 ***
dtrn$expectedpackagesdelivered   0.0000019111   0.0000003201   5.970 0.000000002378196 ***
dtrn$driversworking             -0.0000539415   0.0000292306  -1.845             0.064983 .
dtrn$weatherconditions          -8.9775179164   1.2119691508  -7.407 0.000000000000129 ***
dtrn$pctoversized              103.5264834512  14.5066070956   7.137 0.000000000000957 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 426.24  on 510  degrees of freedom
Residual deviance: 248.23  on 506  degrees of freedom
AIC: 258.23

Number of Fisher Scoring iterations: 7
```

```
> dim(d)
[1] 730  17
> dim(dtrn)
[1] 511  17
> dim(dtst)
[1] 219  17
```

```
> cor(cbind(extrahands=dtrn$extrahands,packages=dtrn$expectedpackagesdelivered,drivers=dtrn
$driversworking,weather=dtrn$weatherconditions,oversz=dtrn$pctoversized))
            extrahands    packages   drivers      weather     oversz
extrahands  1.0000000 0.306516086 0.2068938 -0.278752540 0.22838392
packages    0.3065161 1.000000000 0.7433304  0.006021303 0.02071936
drivers     0.2068938 0.743330385 1.0000000  0.271548971 0.37126448
weather    -0.2787525 0.006021303 0.2715490  1.000000000 0.26527577
oversz      0.2283839 0.020719361 0.3712645  0.265275765 1.00000000
```

Training set predicted vs. actual:

```
> cfm
  trnpred trnactual      count
1       0         0 0.81996086
2       1         0 0.03326810
3       0         1 0.07827789
4       1         1 0.06849315
```

Test set predicted vs. actual:

```
> cfm
  tstpred tstactual      count
1       0         0 0.79908676
2       1         0 0.03652968
3       0         1 0.07305936
4       1         1 0.09132420
```