Ricardo Saucedo
PPHA 30545
Dr. Christopher Clapp
2/9/2021

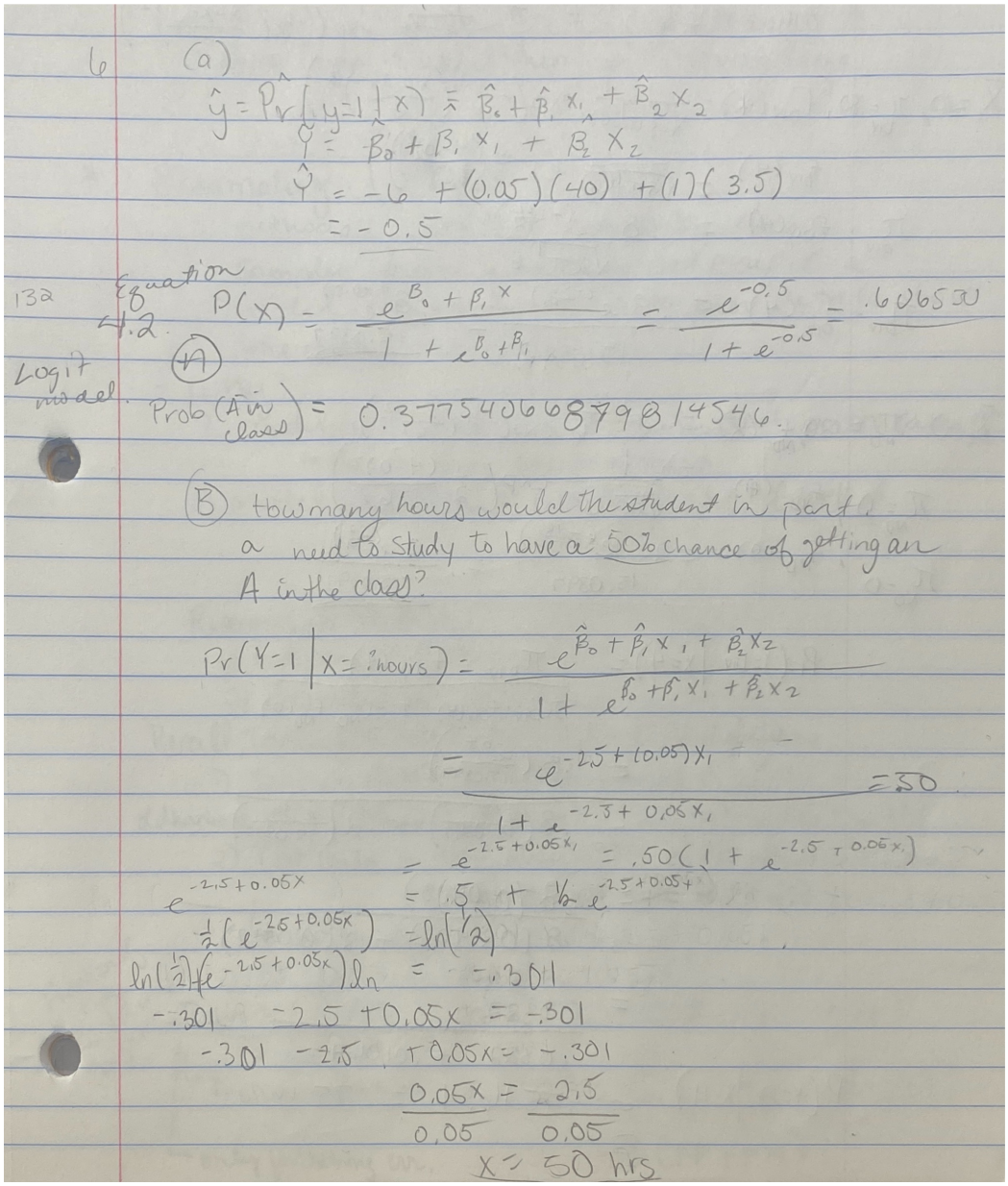
ML Problem Set 2

I worked with Wasil Engel as a partner on this PSet. (UCID: 12231558)

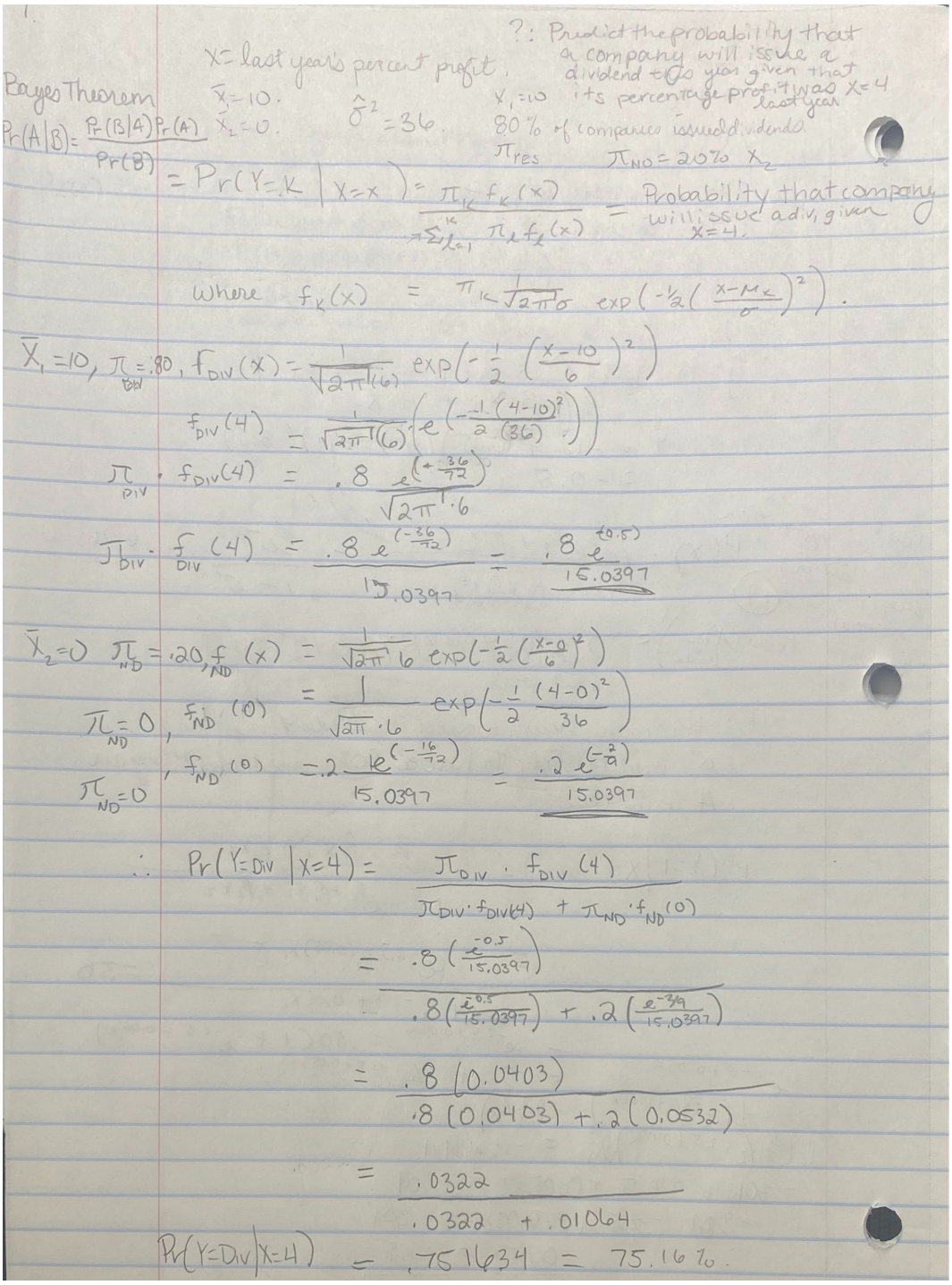
All working code can be located via my GitHub.

Chapter 4 Problems

6.

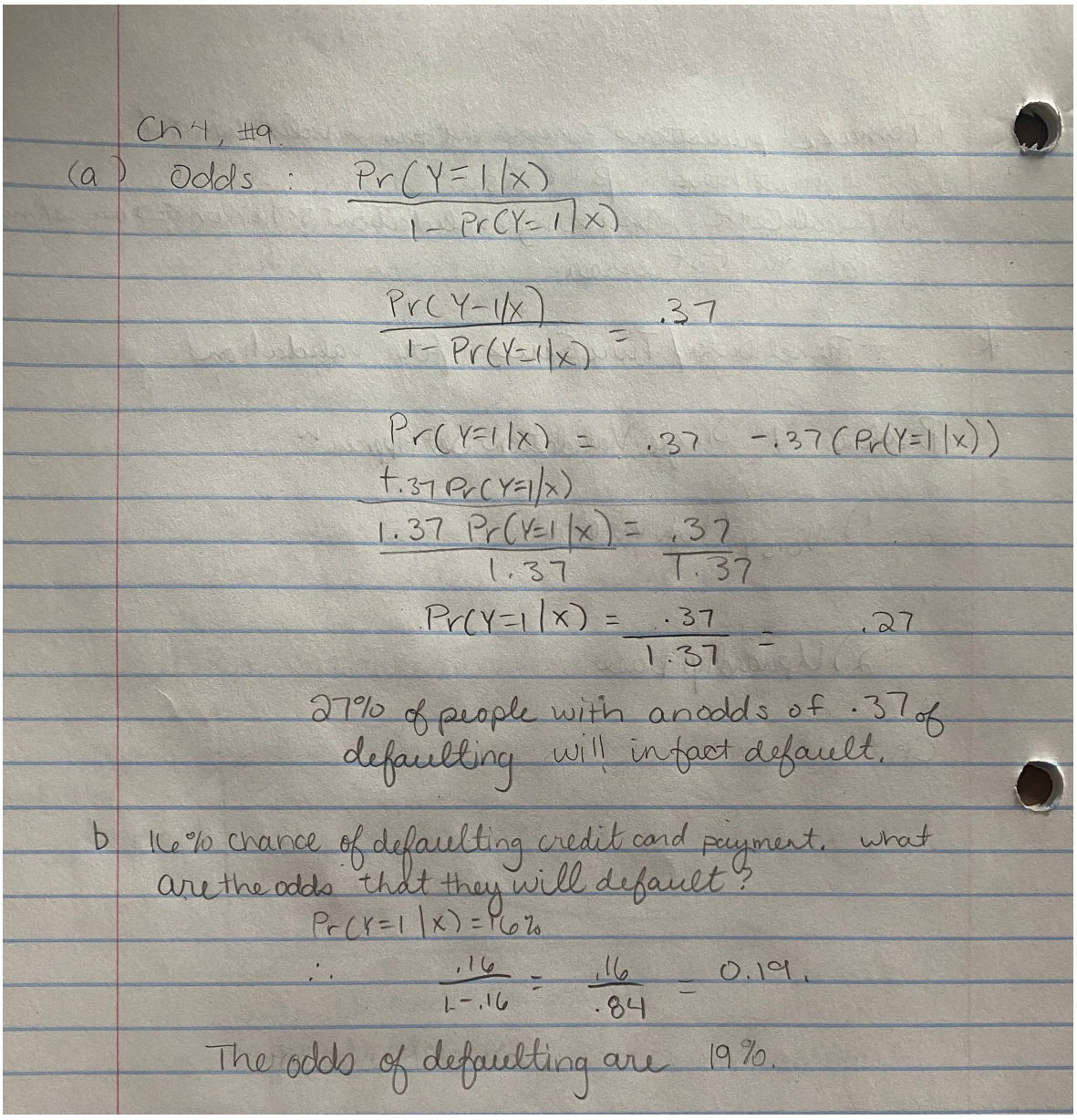


7.



The probability that dividends are given when X = 4 is 75.16%

9.



Ch 4, #9.

(a) Odds : $\frac{Pr(Y=1|x)}{1-Pr(Y=1|x)}$

$\frac{Pr(Y=1|x)}{1-Pr(Y=1|x)} = .37$

$Pr(Y=1|x) = .37 - .37(Pr(Y=1|x))$

$+.37 Pr(Y=1|x)$

$\frac{1.37\ Pr(Y=1|x)}{1.37} = \frac{.37}{1.37}$

$Pr(Y=1|x) = \frac{.37}{1.37} = .27$

27% of people with an odds of .37 of defaulting will in fact default.

b 16% chance of defaulting credit card payment. What are the odds that they will default?

$Pr(Y=1|x) = 16\%$

$\therefore \frac{.16}{1-.16} = \frac{.16}{.84} = 0.19.$

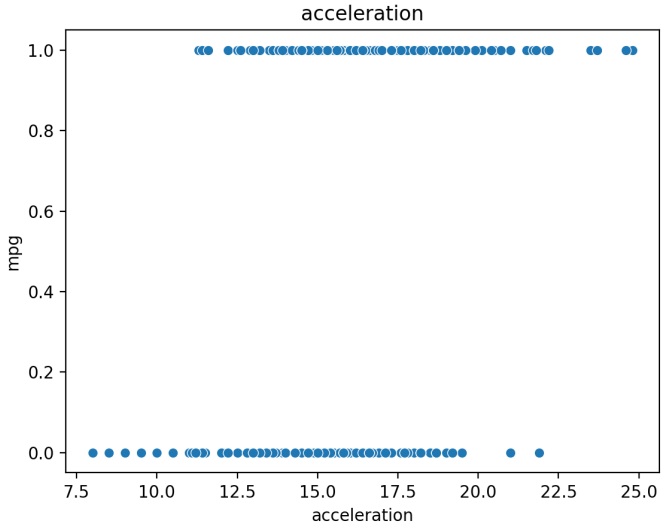
The odds of defaulting are 19%.

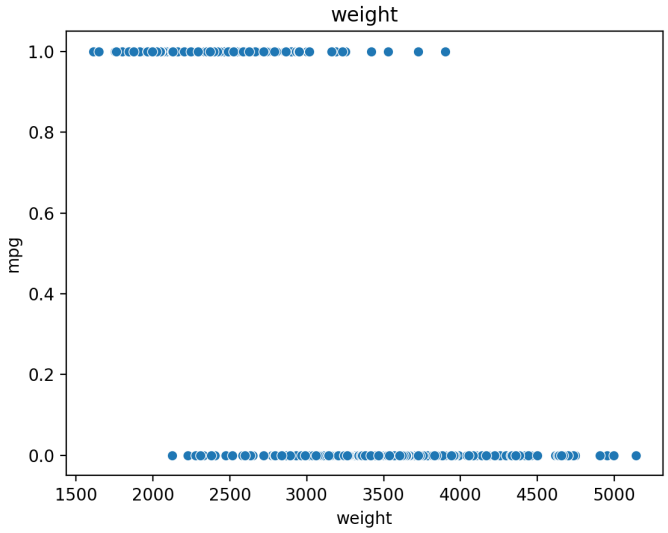
11.

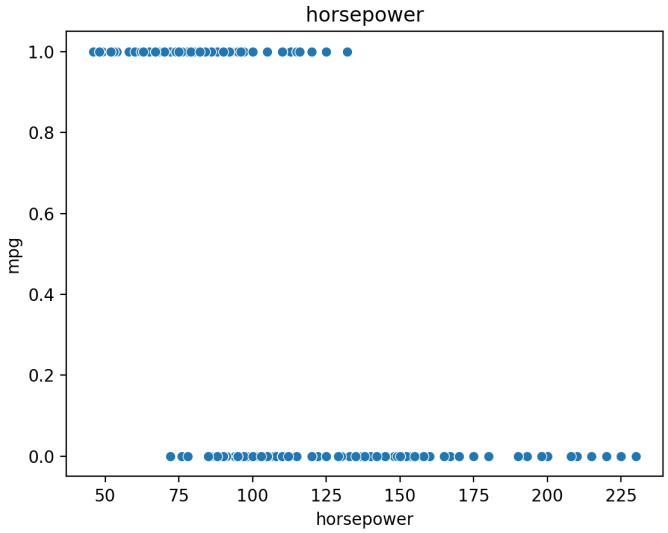
A) Median value is 22.75

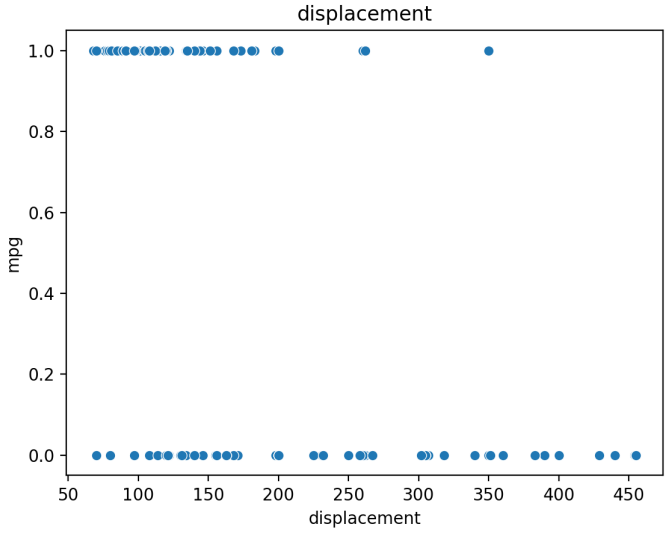
B) Checking correlations, as well as a scatterplot, I feel as if cylinders, displacement, horsepower, weight, and acceleration are good predictors for mpg01. They each have high correlations and intuitively make sense in describing miles per gallon. Even if some

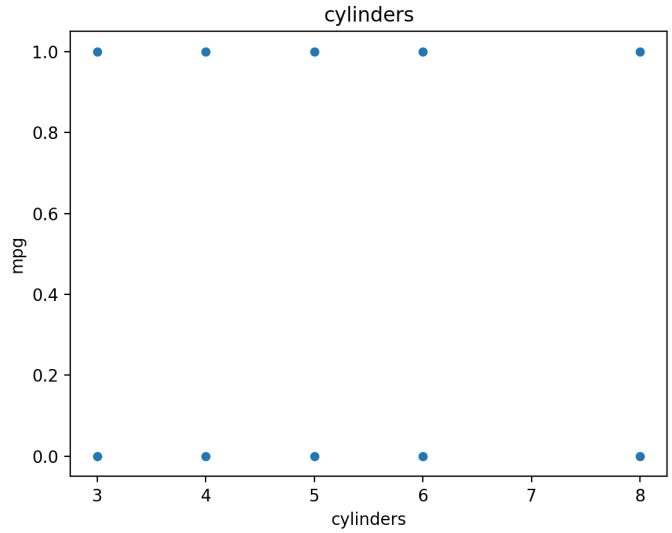
variables are negative, I believe these predictors are still valuable to add into our fitted models. I excluded MPG because this would have quickly overfit our model.





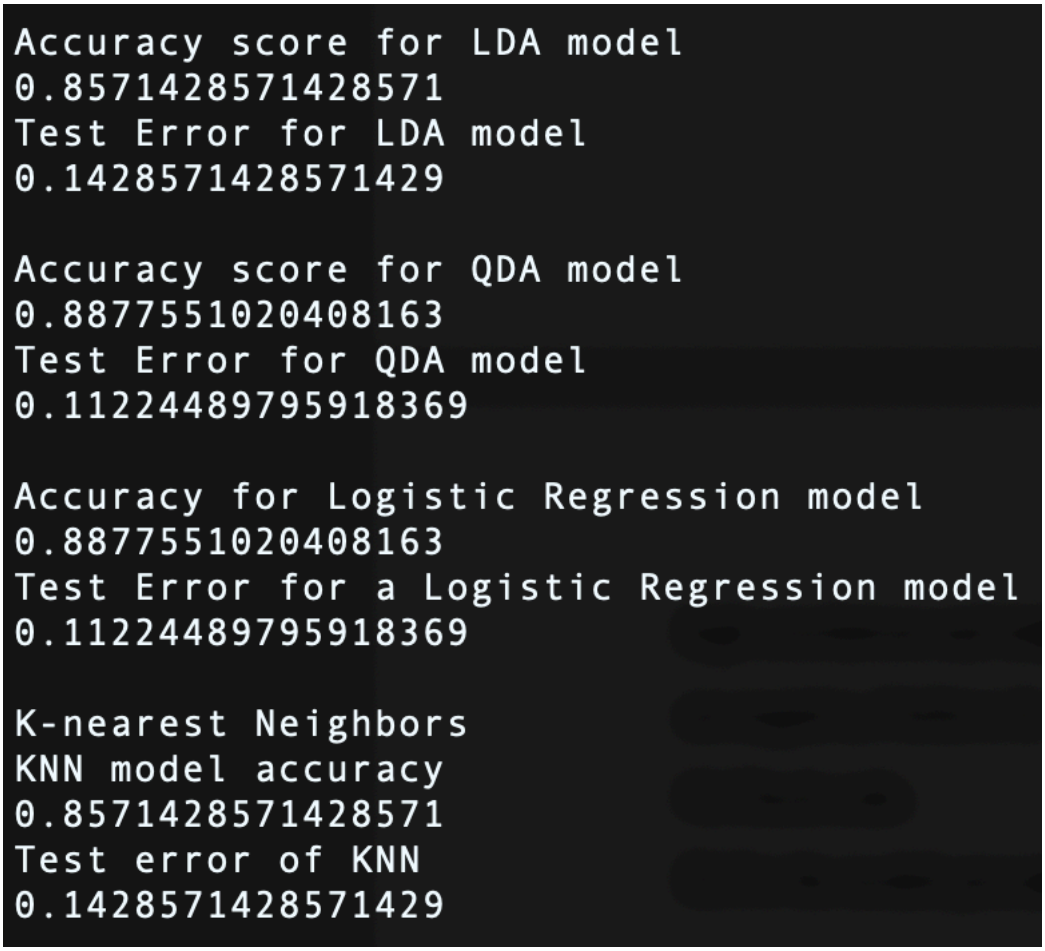

horsepower
mpg
horsepower
1.0
0.8
0.6
0.4
0.2
0.0
50
75
100
125
150
175
200
225


displacement
mpg
displacement
1.0
0.8
0.6
0.4
0.2
0.0
50
100
150
200
250
300
350
400
450



C) Check code.

D, E, F, G) For the models proceeding forward, I decided to go with cylinders, displacement, horsepower, weight, and acceleration.



Floating between 4 and 5, I decided to go with $K = 5$. This gave me the lowest test error, and the highest accuracy score; in reflection, this is the same result we see with our LDA model.

Chapter 5:

5)

```
Chapter 5


part A
Accuracy score
0.9737
Validation set error
0.02629999999999999

part B
part ii: Multiple logistic regression
Accuracy score
0.968
Validation Set Error
0.03200000000000003
part iv
Validation Set Error
0.03200000000000003

part C
Validation set error with set 1  is:  0.027000000000000024
Validation set error with set 2  is:  0.026000000000000023
Validation set error with set 3  is:  0.03433333333333333

part D
Accuracy score Using Balance, Income, and Student status
0.967
Test Error by using Validation Set
0.03300000000000003
```

C) In this case, I observe a validation set error that is different 100% of the time from the prior question. Using different seed values will produce a different subset of possible training and validation sets. Because we are given different answers each time, this is a strong approach, especially when we converge the results from each split so we can arrive at the model's mean accuracy and test error. With what might have appeared to be a high validation set error in part B could actually be much lower when we subset and rerun our model.

D) When we add the dummy variable for student, it appears as if the test error rate increases. There is probable cause that our model could be interpreting more information whenever we add in a dummy variable, therefore changing our prediction. Which a larger test error, we cannot argue that it is insignificant, however, we can argue that the model incorrectly predicts default status by adding in the dummy variable.