Guided Question Set 12 Solutions

```
library(faraway)
library(tidyverse)
library(ROCR)
Data<-wcgs
##train-test split
set.seed(6021) ##for reproducibility
sample<-sample.int(nrow(Data), floor(.50*nrow(Data)), replace = F)
train<-Data[sample, ] ##training data frame
test<-Data[-sample, ] ##test data frame
```

1)

```
##fit model using training data
result<-glm(chd ~ age + sdp + cigs + dibep, family="binomial", data=train)
summary(result)
```

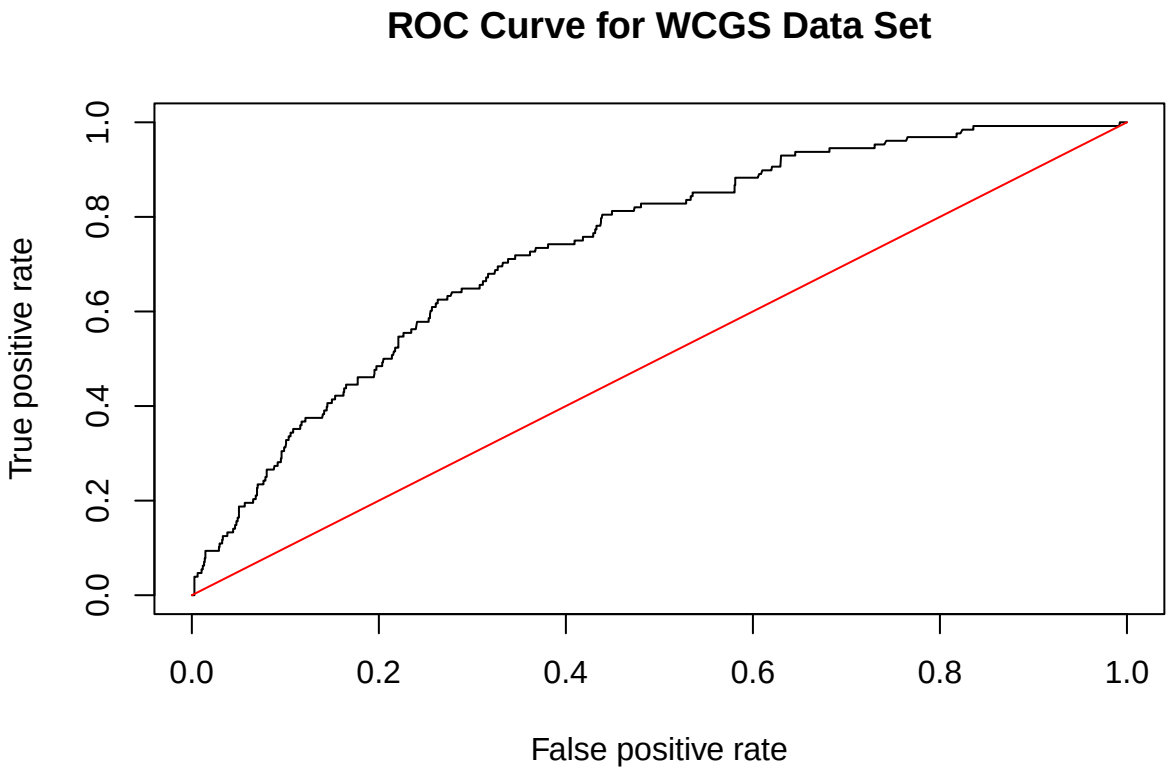
```
##
## Call:
## glm(formula = chd ~ age + sdp + cigs + dibep, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2095  -0.4515  -0.3488  -0.2748   2.6961
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.065578   1.036178  -7.784 7.03e-15 ***
## age          0.060880   0.016560   3.676 0.000237 ***
## sdp          0.020757   0.005595   3.710 0.000207 ***
## cigs         0.020642   0.006035   3.421 0.000625 ***
## dibepB      -0.531792   0.198281  -2.682 0.007318 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 893.04  on 1576  degrees of freedom
## Residual deviance: 838.25  on 1572  degrees of freedom
## AIC: 848.25
##
## Number of Fisher Scoring iterations: 5
```

The estimated coefficients for the quantitative predictors are all positive, which means that increasing age, systolic blood pressure, and the number of cigarettes smoked are associated with higher odds of developing heart disease. The negative coefficient for the indicator associated with type B behaviors means that men with passive behaviors are less likely to develop heart disease, when controlling for the quantitative predictors.

2)

```
preds<-predict(result,newdata=test, type="response")
rates<-prediction(preds, test$chd)
roc_result<-performance(rates,measure="tpr", x.measure="fpr")
plot(roc_result, main="ROC Curve for WCGS Data Set")
lines(x = c(0,1), y = c(0,1), col="red")
```



Since this ROC curve is above the diagonal line, the logistic regression performs better than random guessing.

3)

```
auc<-performance(rates, measure = "auc")
auc@y.values
```

```
## [[1]]
## [1] 0.7371679
```

The AUC of 0.7371679 means the logistic regression performs better than random guessing.

4)

```
table(test$chd, preds>0.5)
```

```
##
##       FALSE
##   no   1449
##   yes   128
```

With a threshold of 0.5, the accuracy is $\frac{1449}{1449+128} = 0.92$, TPR is 0, and the FPR is also 0. Notice that no observation in the test data is predicted to develop heart disease when the threshold is 0.5.

5)

I agree with the classmate. This means that for a threshold of 0.5, the ROC curve is at the bottom left corner, right where the diagonal is.

Also, this confusions means that regardless of the true status of the test observations, they are all classified as not developing heart disease.

6)

We may be more concerned with correctly identifying observations who have heart disease, so we want to increase the TPR. The cost of missing these observations would be that they cannot be treated for heart disease.

We may be willing to accept a corresponding increase in FPR: more observations who do not have heart disease get incorrectly classified as having heart disease. The cost of this mistake could be less consequential.

Thus, we want to lower the threshold, to make it easier to classify an observation as having heart disease.

7)

Consider a threshold of 0.08

```
table(test$chd, preds>0.08)
```

```
##
##       FALSE TRUE
##   no    940  509
##   yes    36   92
```

We now have an accuracy of $\frac{940+92}{940+509+36+93} = 0.65$, a TPR of $\frac{92}{36+92} = 0.72$ and a FPR of $\frac{509}{940+509} = 0.35$. As noted earlier, we increase the TPR at the expense of the FPR.

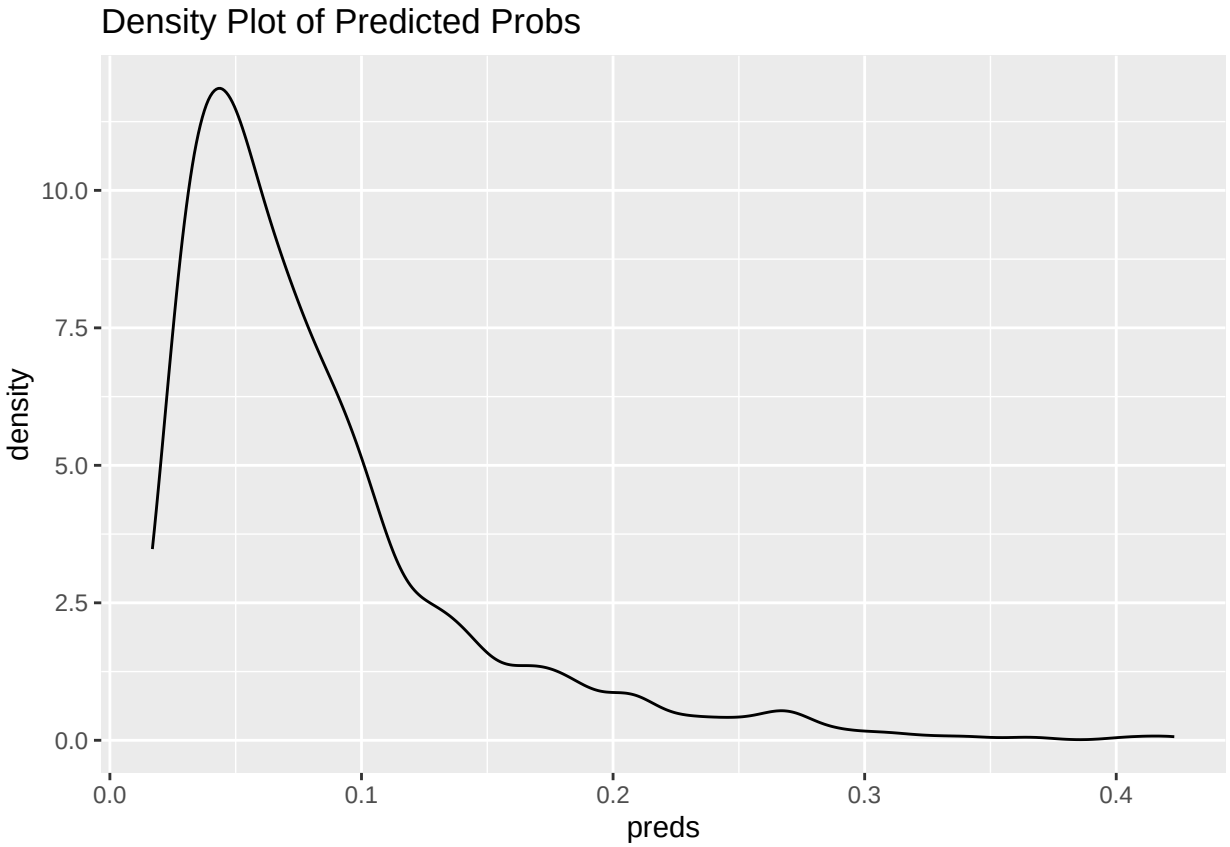
Notice that the accuracy has decreased, even though we are doing better than random guessing now.

8)

From part 4, we notice that 1449 men did not develop heart disease, while only 128 men did. So developing heart disease is a rare event, $\frac{128}{1449+128} = 0.08$. Using a threshold of 0.5 to classify an observation as having heart disease may be unrealistic in this context, since we are modeling a rare event.

If we create a density plot of the predicted probabilities of developing heart disease in the test data, notice that most of the observations have probabilities less than 0.1. The largest probability is slightly more than 0.4. So none of the predictors result in high estimated probabilites for any observation.

```
test<-data.frame(test,preds)
ggplot(test,aes(x=preds))+
  geom_density()+
  labs(title="Density Plot of Predicted Probs")
```



Using a threshold of 0.5 will minimize the overall error rate. Is our consideration overall error rate, or reducing the false positive rate or reducing the false negative rate? Consultation with a subject matter expert will be needed.