**Data Analytics – I415/H515**

**FINAL EXAM**

**Out: Sunday, December 12, 2021, 11:59 pm**

**Due: Sunday, December 19, 2021, 11:59 pm**

*Instructions:*

The exam is open notes and open book. You **are not allowed** to consult any other individual, be they enrolled in the class or not, about the exam.

In writing your solutions, make them complete yet concise. Mysterious answers unsupported by explanation or derivation will not earn full marks, while a wrong answer accompanied by derivation may earn partial credit.

**If at any point you use R to aid in the solution to a problem, include your code AND its output. Also, be sure to use any provided R code in the FinalExam.R file in order to not miss unnecessary points.**

**Any clarification questions you have while working on the exam must be posted in the Final Exam Discussion Board so that everyone has the same information.**

There are four problems on this exam. Each problem is worth 25 points. The first three problems are broken down further into each part of the problem. The fourth problem provides you with two options to complete and will be graded holistically.

1. [25 pts] A key performance for hospitals is the *30-day unplanned readmission rate*—the proportion of patients discharged from the hospital who had an unplanned readmission within 30 days. Programs like the Hospital Readmissions Reduction Program (HRRP) apply penalties (up to a 3% reduction in payments) to underperforming U.S. hospitals—resulting in withheld payments in excess of $500 million in 2018.

Hospitals can employ some low-cost strategies to reduce unplanned readmissions, such as confirming patient follow-up plans prior to discharge and asking patients to verbally repeat their treatment directions. However, other approaches are more involved and costly. One example is to arrange “telehealth” interventions, in which health care providers contact patients routinely after discharge. Given the cost of these interventions, they are only appropriate for patients at elevated risk of readmission.

You are working for a mid-sized hospital in the northeast United States, and are tasked to assess the impact of telehealth interventions on diabetic patients—with the ultimate goal of reducing the 30-day read- mission rate. The intervention will cost approximately $1,200 per patient. Clearly, it must be limited in scope, and a key component of your strategy will be targeting the “right” patients.

Unfortunately, your hospital does not document 30-day readmissions, as this requires significant follow-up with discharged patients. You will thus use a publicly-available dataset to study readmission risk. The dataset includes over 100,000 hospital discharges of over 70,000 diabetic patients from 130 hospitals across the United States during the period 1999–2008. All patients were hospital inpatients for 1–14 days, and received both lab tests and medications while in the hospital. The 130 hospitals represented in the dataset vary in size and location: 58 are in the northeast United States and 78 are mid-sized (100–499 beds).

The dataset is provided in the “readmission.csv” file. It contains the following variables:

* **readmission**: 1 if the patient had an unplanned readmission within 30 days, 0 otherwise
* **Patient characteristics**: race, gender, and age capture demographic information.
* **Recent medical system use**: The variables numberOutpatient, numberEmergency, and numberInpatient capture the number of times the patient used the medical system in the last year.
* **Diabetic treatments**: A number of variables capture the patient’s diabetic treatments: acarbose, chlorpropamide, glimepiride, glipizide, glyburide, glyburide.metformin, insulin, metformin, nateglinide, pioglitazone, repaglinide, and rosiglitazone.
* **Admission information**: The variables admissionType and admissionSource contain information about how the patient was admitted to the hospital. The variable numberDiagnoses captures the number of diagnoses the patient had recorded for their admission. There are also a number of variables that indicate whether a patient was diagnosed with various conditions when admitted: diagAcuteKidneyFailure, diagAnemia, diagAsthma, diagAthlerosclerosis, diagBronchitis, diagCardiacDysrhythmia, diagCardiomyopathy, diagCellulitis, diagCKD, diagCOPD, diagDyspnea, diagHeartFailure, diagHypertension, diagHypertensiveCKD,diagIschemicHeartDisease, diagMyocardialInfarction, diagOsteoarthritis, diagPneumonia, and diagSkinUlcer.
* **Treatment information**: timeInHospital is the number of days the patient was in the hospital, and numLabProcedures, numNonLabProcedures, and numMedications capture the amount of care the patient received in the hospital.

1. Open the data file “readmission.csv”in R. Perform some exploratory data analysis on the full data set and report two interesting insights you gained from your analysis. [2 pts]

*YOUR SOLUTION:*

*You can see from the exploratory data analysis I did below that this study had majority white male participates.*

*readmission %>%*

*group\_by(gender) %>%*

*summarise(gender\_count=n())*

*1 Female 54708*

*2 Male 47055*

*3 NA 3*

*readmission %>%*

*group\_by(race) %>%*

*summarise(gender\_count=n())*

*1 AfricanAmerican 19210*

*2 Asian 641*

*3 Caucasian 76099*

*4 Hispanic 2037*

*5 Other 1506*

*6 NA 2273*

1. Based on conversations with the hospital’s management, you estimate the cost of a **30-day unplanned readmission at $35,000**. From published information at a similar institution, you estimate that tele- health interventions will reduce the incidence of 30-day unplanned readmissions in the treated popuation by 25%. Given the cost of $1,200 per intervention, what are:
   * the “loss” of a false positive, as compared to a true negative; and
   * the “loss” of a false negative, as compared to a true positive?

Define the loss matrix for your CART model. [2 pts]

*YOUR SOLUTION:*

*cost.matrix <-matrix(c(0, 1200, 35000, 0),2,2,byrow=T)*

*L\_FP <- 1200*

*cost <- (35000 \* .25)*

*cost1<- 35000 - cost - 1200*

*L\_FN <- cost1*

*loss.matrix <- matrix(c(0,L\_FP,L\_FN,0),2,2,byrow=T)*

*loss.matrix*

*> cost.matrix*

*[,1] [,2]*

*[1,] 0 1200*

*[2,] 35000 0*

*> loss.matrix*

*[,1] [,2]*

*[1,] 0 1200*

*[2,] 25050 0*

*For every 1200 false positives that are predicted in the loss matrix, its worth 25,050 in false negatives.*

*For every 1200 false negatives, there is predicted to be 35,000 true positives.*

1. Fit a CART model using a cp parameter of 0.001 and the loss matrix defined in Question b. Include an image of your tree. [8 pts]

*YOUR SOLUTION:*

*set.seed(121)*

*tree <- rpart(readmission ~ ., data = readm.train, parms =list(loss=loss.matrix), control = rpart.control(cp = 0.001))*

*prp(tree)*

*Diagram

Description automatically generated*

1. Assess the model’s predictive performance using the test set. What is the accuracy, true positive rate and false positive rate?
   * Contrast the decisions resulting from your model and those resulting from current practice (under which no patient is subject to a telehealth intervention). Provide summary statistics to explain how the decisions differ, and discuss the costs and benefits of each approach. Make sure to compare the total monetary costs of patient readmission. [7 pts]

*YOUR SOLUTION:*

*pred.cart = predict(tree, newdata=readm.test)*

*confusion.matrix <- table(readm.test$readmission , pred.cart)*

*number.interventions <- confusion.matrix[1,2]+confusion.matrix[2,2]*

*prevented.readmissions <- confusion.matrix[2,2]\*.25*

*> number.interventions*

*[1] 8238*

*> prevented.readmissions*

*[1] 196.75*

*accuracy <- sum(diag(confusion.matrix)) / sum(confusion.matrix)*

*TPR <- confusion.matrix[2,2]/sum(confusion.matrix[2,])*

*FPR <- confusion.matrix[1,2]/sum(confusion.matrix[1,])*

*> accuracy*

*[1] 0.3506938*

*` > TPR*

*[1] 0.2871215*

*> FPR*

*[1] 0.3282379*

*baseline.accuracy <- sum(!readm.test$readmission)/nrow(readm.test)*

*baseline.cost <- (confusion.matrix[2,1]+confusion.matrix[2,2])\*35000*

*absolute.savings <- (baseline.cost-accuracy)/baseline.cost*

*relative.savings <- (baseline.cost-cost)/baseline.cost*

*> baseline.accuracy*

*[1] 0.8922605*

*> baseline.cost*

*[1] 48615000*

*> absolute.savings*

*[1] 1*

*> relative.savings*

*[1] 0.99982*

*According to the data above you can see that the FPR ,TPR and Accuracy are all fairly low. When it comes to the baseline cost right now it is very high, accounting to the data, each time we try to prevent a readmission we are saving nearly $200, this is about a fraction of each night if they were to stay. Absolute and relative savings are very high and would be what we should be looking over to have patient make telehealth appointments Today’s current practice is spending 35000 for every 30 days a patients stay. The data is telling us that we would reduce this by a fraction of the cost per day.*

1. Can cross validation improve your model or is a cp of .001 optimal? [3 pts]

*YOUR SOLUTION:*

*The cp in this case is optimal because the predictor is a binary response, it wouldn’t give good results if you were to use a cross validation.*

*e)* So far, you have selected the subset of patients that maximizes total net value, given the cost of the telehealth intervention and the willingness to pay for a prevented readmission. However, in practice you might have to adjust your model to try to directly improve the true positive rate (TPR) and/or the false positive rate (FPR).

1. How would you modify the loss matrix from part b) to obtain a CART model with a higher TPR than the one in part c)? [1.5 pts]

*YOUR SOLUTION:*

*To improve the loss matrix I would try to better understand the cost matrix, and potentially add another value that would need to be measured in the cost matrix.*

1. How would you modify the loss matrix from part b) to obtain a CART model with a lower FPR than the one in part c)? [1.5 pts]

*YOUR SOLUTION:*

*To lower the FPR, I again try to factor in more variables that are associated with the cost in the cost matrix.*

1. [25pts] For historical reasons the US has a system of taxing homeowners to fund a large fraction of local infrastructure such as local primary, middle and high schools, town and county administrations, town and county roads, …. The tax, called “property tax”, is based on an assessment (estimation, determination) of the value of each residence (home) and the lot (land) that belongs to it. Because the assessments become outdated after a few years, towns have to hire assessors and update the assessments every so often.

There are of course several factors that play a role in assessing the value of a property: square feet of livable area, size of the lot (land), quality and condition of the building, desirability of the area, … We will only consider square feet of livable area. The following exercise can be used to check for any property whether its assessed value is in line or out of line with other properties of similar size in terms of livable area. This dataset is called “Residential\_Property\_Assessments.csv.”

* 1. Show a scatterplot of *Assessment* against *Livable Area* (i.e., *Assessment* is the y variable, and *Livable Area* is the x variable). Add a main title along with axis labels. [3pts]

*YOUR PLOT:*

*Chart, scatter chart

Description automatically generated*

* 1. Based on the scatterplot, is the association approximately linear? [1pt]

*YOUR ANSWER:*

*Yes, the scatterplot has a strong positive correlation.*

* 1. Use R to find the equation for the regression line with *Assessment* as the dependent variable and *Living Area* as the independent variable. [2pt]

*YOUR SOLUTION:*

*equ <- lm(Assessment~., property)*

*equ*

*Assessment = 168518.6 + 180.8 \* Living Area*

* 1. Interpret the slope. Initially, do so first formally according to the formulation from class. Then, give an informal interpretation in terms of the average value of a square foot of livable area. [4pts]

*YOUR SOLUTION:*

*Formally: A different in Living Area by 1 unit is approximately a difference of 180.8 in beta1.*

*Informally: For every single square foot in Living Area, there is 180.8 added to the total Assessment.*

* 1. Interpret the intercept, first formally [1pt], then explain why this is not meaningful [2pts]. [3 pts total]

*YOUR SOLUTION:*

*Formally: The intercept holds the value of y when x equal zero.*

*This is not meaningful because all numbers in the set are positive.*

* 1. What fraction of variation in *Assessment* is accounted for by *Livable Area*? Report the relevant quantity from the R output [2pt]. Finally, what fraction of variation is accounted for by other factors besides livable area, such as differences in lot size, condition and quality of the building, viability of the area…? [2pts]. [4pts total]

*YOUR SOLUTION:*

*The* *fraction of variation in this linear regression r-squared is 0.7893.*

*The fraction of variation that is account for by other factors is the difference of the R-squared above, in this case it would be .22.*

* 1. Based on the fitted equation, what can you say about the predicted price for a residence with 2,500 sqft of livable area? Use R to calculate this and show your code [4 pts].

*YOUR SOLUTION:*

*new <- data.frame( Livable.Area = c(2500))*

*this <- predict(equ, newdata = new, interval = "confidence")*

*fit lwr upr*

*620569.1 617430.3 623707.9*

* 1. In general, how much uncertainty is left over after making our predictions using the regression line? Quote the relevant standard deviation from R. Compare this quantity to the variability left over if we had instead simply predicted the average of Assessment for all individuals, rather than using information about Livable Area to improve our prediction. [4 pts]

*YOUR SOLUTION:*

*The standard error in this linear regression is 5.894.*

*When looking at the lwr and upr above, you can see that they are fairly close to each other, which leave little room for uncertainty when it comes to this prediction.*

*If we were to simply predict the average off of the Assessment, it wouldn’t capture the data or take into account outliers as well or the size of the property which is crucial to trying to predict.*

1. [25pts] We have data on n=78 individuals who were diagnosed with breast cancer, had the tumor surgically removed, and had no identifiable trace of the disease in their bodies after the surgery. Five years later, for each of these individuals we have information on whether or not their cancer came back over the course of the five years versus whether or not they stayed in remission. The data set in “breastcancer.csv” contains 7 clinical covariates, such age, and tumor diameter, along with 4,348 genetic indicators of prevalence of certain genes. The goal is to see whether or not reoccurrence of breast cancer after surgery is predictable based on both clinical and genetic information. If possible, clinicians would like to identify those at high risk for recurrence and recommend adjuvant therapy post-surgery, such as chemotherapy, as a means of reducing the risk of recurrence. The data set contains 34 individuals whose cancer did come back (labeled as response=1), while the remaining 44 did not (response = 0). In terms of data structure, this data set differs from those we’ve previously studied in that the number of covariates is much larger than the actual number of patients in the study. We will now see that this introduces major complications which need to be handled with care.

Before proceeding, execute the provided code to partition the data set into training set and test set. The training set is of size 50, and the test set has 28 individuals.

* 1. Our first instinct would likely be to fit a multiple logistic regression model using glm()on the training set. Try this out in R. You’ll notice that the output of the summary command looks bizarre. How many slope coefficients do NOT have the value NA? How does this compare the size of the training set? The function is.na()will be helpful here. [3 pts]

*YOUR SOLUTION:*

*mult.logit.bc = glm(response~., data = bc.train, family = "binomial")*

*summary(mult.logit.bc$df.null)*

*There are 49 slope coefficients without NA values.*

*Compared to the size of the data set this is very low, there are over 4k attributes in this data set.*

* 1. Nonetheless, proceed using the predict() function in R to predict the responses in the training set based on our model. Use the command round(..., digits = 10) to round the estimated probabilities to 10 decimal places. Looking at your training set, what does the distribution for the predicted probabilities look like? [3 pts]

*YOUR SOLUTION:*

*pred.mult.train = predict(mult.logit.bc, bc.train, type ="response")*

*round(pred.mult.train, digits = 10)*

*8 9 10 12 13 14 15 16 17 18 19 21 22 23 24 27 28 29 30 32*

*0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0*

*33 35 36 39 40 42 43 44 45 46 47 50 52 53 55 57 58 60 61 63*

*0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1*

*64 65 66 67 68 69 70 73 74 78*

1. *1 1 1 1 1 1 1 1 1*

*As you can see by the numbers, the graph is mostly skewed to the left, with all the numbers being towards the end of the dataset.*

* 1. Construct the in-sample confusion matrix based on this algorithm. What are the False Positive Rate and True Positive Rate in-sample? And what would the AUC be? [4 pts]

*YOUR SOLUTION:*

*Below you can see the False Positive and True Positive Rates for in-sample, the AUC for this is 1 because the data comes from the same set.*

*truePos = confusion.matrix[2,2]/sum(confusion.matrix[2,])*

*truePos*

*# 0*

*falsePos = confusion.matrix[1,2]/sum(confusion.matrix[1,])*

*falsePos*

*# 0.03571429*

*rocr.pred = prediction(pred.mult.train, bc.train$response)*

*roc.perf = plot(performance(rocr.pred, "tpr", "fpr"))*

*AUC = as.numeric(performance(rocr.pred, "auc")@y.values)*

*AUC*

*# 1*

* 1. A system of n equations with n unknown parameters will always have at least one solution, while a system of n equations with more than n unknown parameters will have infinitely many solutions. In light of this, why is it that we were able to attain in-sample error rates observed in part c)? [3pts]

*YOUR SOLUTION:*

*We are unable to attain in-sample error rate because of how the data is set up in a binary format. Its also hard to a system of equation when you have the numbers set up in a matrix format with nothing but 0’s on one half.*

Situations where the number of covariates are larger than the number of observation provide further motivation for using regularized regression. We clearly would like to include all of the potentially relevant covariate information, but mathematical deficiencies of the conventional solution render this difficult. The regularization provides automatic variable selection, while making the observed fit less susceptible to overfitting.

In the code you see we have given you code for running the “LASSO” version of logistic regression while cross-validating the tuning parameter, which is facilitated by the cv.glmnet function.

* 1. How many nonzero slope coefficients are there in the resulting penalized logistic regression? [3pts]

*YOUR SOLUTION:*

*There are 19 nonzero slopes*

*this <- coef(lasso.logit, s = "lambda.min")*

*summary(this)*

*i j x*

*1 1 1 -0.33103484*

*2 2 1 -0.26482886*

*3 6 1 0.10470011*

*4 35 1 -0.02281052*

*5 292 1 0.37021982*

*6 627 1 -0.01089136*

*7 1262 1 0.15992673*

*8 1798 1 -0.07898490*

*9 2024 1 0.42674656*

*10 2218 1 -0.05638389*

*11 2228 1 -0.02246715*

*12 2430 1 0.16779682*

*13 2665 1 -0.19757751*

*14 2701 1 -0.03665294*

*15 3401 1 0.58442708*

*16 3742 1 0.03238161*

*17 3798 1 0.19128971*

*18 3939 1 -0.15816417*

*19 3974 1 0.07111450*

* 1. Compute the in-sample predictions from the regularized logistic regression. Present a histogram of the predictions, and compare this to what we saw in part b). [4pts]

*YOUR SOLUTION:*

*As you can see from the graphs below the data looks very skewed to opposite ends, compared to the prediction we just ran that is more distributed.*

*pred.lasso.train = as.vector(predict(lasso.logit, newx = x.train, s = "lambda.min", type = "response"))*

*hist(pred.mult.train, breaks=5, main="3B Predict")*

*hist(pred.lasso.train, breaks=5, main="3F Predict")*

Chart, histogram

Description automatically generated

Chart, histogram, scatter chart

Description automatically generated

* 1. Compute the out-of-sample AUC for the two we’ve developed here: multiple logistic regression and the regularized logistic regression. What do you find? Which one does best out of sample? And are there any algorithms that seem to perform worse than random coin flips? [5pts]

*YOUR SOLUTION:*

*pred.mult.test = predict(mult.logit.bc, newdata = bc.test, type = "response")*

*pred.lasso.test = as.vector(predict(lasso.logit, newx = x.test, s = "lambda.min", type = "response"))*

*prediction.mult.test= prediction(pred.mult.test, bc.test$response)*

*perf.mult.test = performance(prediction.mult.test, measure = "tpr", x.measure = "fpr")*

*prediction.lasso.test= prediction(pred.lasso.test,bc.test$response)*

*perf.lasso.test = performance(prediction.lasso.test, measure = "tpr", x.measure = "fpr")*

*roc.analysis.lasso = roc(bc.test$response ~perf.lasso.test, bc.test)*

*roc.analysis.lasso$response*

*roc.analysis.mult = roc(bc.test$response~pred.mult.test, bc.test)*

*roc.analysis.mult$response*

*plot(perf.mult.test, xlab = "False Positive Rate", ylab = "True Positive Rate", main = "Out-of-Sample ROC Comparison")*

*plot(perf.lasso.test, add= T, col = "orange", lty = 4, lwd = 2)*

*legend("bottomright", c("Multiple Logistic", "L1 Penalized"), lty = c(1, 4), col = c("black", "orange"), bty="n")*

*You can see by the graph that the the regularized logistic regression does better out of sample. Its closer to a corner then the black line.*

*Chart

Description automatically generated*

1. For this question, you may choose from two unstructured options. Choose only one option and do not submit answers for both. You need to complete steps 2-6 of the basic data analytics process outlined below for the option you choose. This question will be graded holistically and is worth 25 points

**Basic Data Analytics Process**

1. ~~Generate a question and collect data.~~
2. Visualize, prepare data, and understand your data.
   1. If needed, clean and normalize the data. Also, consider removing redundant or known correlated variables.
   2. Provide insights you gain from data exploration.
3. Choose a model or models we covered in this class.
   1. Explain why you chose the model or models to answer the primary question.
   2. Explain whether you value accurate predictions above all else or more interpretable and actionable model(s).
4. Train and tune the model(s) on the training set.
5. Assess the model(s) on the test set.
6. Provide recommendations and analysis of your results.

Option 1

After a recent trip to a winery with some of my friends, who happen to be chemists, we developed a method that would enable us to **alter physicochemical** attributes of red wine to create the world’s highest quality wine. We hypothesized that we could even alter bagged wine to be delicious. Unfortunately, we did not take the time to figure out exactly what qualities we should alter to what levels. Fortunately, we believe you have the skills necessary to do so.

The file “redwine.csv” is related to red variants wine. The data only includes physicochemical (inputs) and sensory (the output) variables. There is no data about grape types, wine brand, wine selling price, etc.

Would you please help us understand what variables matter most and how we should adjust our wine to create the highest quality wine in all the land? [25pts if chosen]

*YOUR SOLUTION:*

*wine = read.csv("redwine.csv")*

*split = createDataPartition(wine$fixed.acidity, p = 0.70, list = FALSE)*

*wine.train <- wine[split,]*

*wine.test <- wine[-split,]*

*x.train = wine.train[,-1]*

*x.test = wine.test[,-1]*

*y.train = wine.train[,1]*

*y.test = wine.test[,1]*

*plot(wine)*

Diagram

Description automatically generated

*You can see in the graph that they’re only a handful of graphs that you can see have a clear correlation.*

*You can see by the code below that I ran multiple linear regression models on all attributes in the dataset to see which predictor would produces the best r-squared. I believed that running a linear regression model on this data would be most effective because of the multiple continuous variables. I wanted to get the most accurate model so that we’re able to predict what factors would make the best wine.*

*I first ran multiple models and chose the top two that produces the best r-squared factor. I then took out the predictors that weren’t significant in each of the models, and created new models for each model I chose. In this case the models that did the best were the ones with the predictors fixed.acidity and density.*

*I was able to then run these two models on my test set, where I realized the data may have been overfitted because the same attributes that were captured as significant in the training set weren’t significant in the test set. I then re-ran both models with cross validation and came to the conclusion that the density model wouldn’t work well to predict because I came out with a negative r-squared when I tested the cross validation density model on the test set.*

*In conclusion, I found that the factors that don’t have as much significance in this dataset are volatile.acidity, free.sulfur.dioxide , quality. The rest of the factors hold the most weight when trying to predict which wine would be the best.*

*fix <- lm(fixed.acidity~., wine.train)*

*summary(fix)*

*vola <- lm(volatile.acidity~., wine.train)*

*summary(vola)*

*cit <- lm(citric.acid~., wine.train)*

*summary(cit)*

*sugar <- lm(residual.sugar~., wine.train)*

*summary(sugar)*

*chlor <- lm(chlorides~., wine.train)*

*summary(chlor)*

*free <- lm(free.sulfur.dioxide~., wine.train)*

*summary(free)*

*total <- lm(total.sulfur.dioxide~., wine.train)*

*summary(total)*

*den <- lm(density~., wine.train)*

*summary(den)*

*ph <- lm(pH~., wine.train)*

*summary(ph)*

*sul <- lm(sulphates~., wine.train)*

*summary(sul)*

*alch <- lm(alcohol~., wine.train)*

*summary(alch)*

*qual <- lm(quality ~., wine.train)*

*summary(qual)*

*den1 <- lm(density~.-citric.acid -quality, wine.train)*

*fix <- lm(fixed.acidity~., wine.train)*

*den.test <- lm(density~.-citric.acid -quality, wine.test)*

*fix.test <- lm(fixed.acidity~.- quality , wine.train)*

*set.seed(123)*

*train.control <- trainControl(method = "cv", number = 5)*

*model1 <- lm(fixed.acidity~., wine.train)*

*summary(model1)*

*model1 <- train(fixed.acidity ~.-quality, data = wine.train, method = "lm",*

*trControl = train.control)*

*model2 <- train(fixed.acidity ~.-volatile.acidity- free.sulfur.dioxide -quality , data = wine.train, method = "lm", trControl = train.control)*

*pred = predict(model1, newdata=wine.test)*

*r2 = 1-sum((y.test - pred)^2)/sum((y.test - mean(y.train))^2)*

*r2*

*#[1] 0.867495*

*pred = predict(model2, newdata=wine.test)*

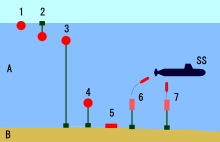
*r2 = 1-sum((y.test - pred)^2)/sum((y.test - mean(y.train))^2)*

*r2*

*#0.8676707*

Option 2

As you are aware, I have spent some of my career beneath the waves on a submarine. Hazards come with the job, but what you may not know, is that that underwater mines are real and can be a threat to submarines. Rocks, on the other hand, are relatively harmless to pass over. (<https://en.wikipedia.org/wiki/Naval_mine>)



I would like to help my friends still serving on submarines by developing a model that could predict whether or not something is a mine or a rock from sonar data. They could then ignore the rocks and avoid the mines. Luckily, I found an amazing dataset on the internet to help me do just that.

The file "rockormine.csv" contains 111 patterns obtained by bouncing sonar signals off a “mine” at various angles and under various conditions and 97 patterns obtained from rocks under similar conditions. The transmitted sonar signal is a frequency-modulated chirp, rising in frequency. The data set contains signals obtained from a variety of different aspect angles, spanning 90 degrees for the cylinder and 180 degrees for the rock.  
  
Each pattern is a set of 60 numbers in the range 0.0 to 1.0. Each number represents the energy within a particular frequency band, integrated over a certain period of time. The integration aperture for higher frequencies occur later in time, since these frequencies are transmitted later during the chirp.  
  
The label associated with each record contains the letter "R" if the object is a rock and "M" if it is a mine. The numbers in the labels are in increasing order of aspect angle, but they do not encode the angle directly.

Please help me keep my friends safe by developing a model that can best predict whether an object is a rock or a mine. While model accuracy is of upmost importance, do not penalize false positives or false negatives more than the other. Doing so could train your model to avoid every rock in the ocean, making it impossible for the submarine to even pull out of port. [25pts if chosen]

*YOUR SOLUTION: …*