1. You are tasked with designing a movie recommendation model for Netflix. You build a logistic regression model to predict the probability that a viewer likes a movie. In the test data set of 10 movies, the model delivers the following probabilities (with the class label = 1 if the viewer likes the movie, and 0 if she does not.)

Movie id	Pred. Probability	True class
1	0.91	1
2	0.89	1
3	0.84	0
4	0.82	1
5	0.79	1
6	0.73	0
7	0.62	1
8	0.54	0
9	0.49	0
10	0.48	0

   a. Viewers get upset if recommended a movie that they don't like and so you set a high threshold of probability 0.8 or higher to recommend a movie. Write down the confusion matrix for the above data using this threshold.

   b. What is the accuracy of the model in the test data?

   c. What is the precision in the test data? The recall?

   d. Compute the true positive rate (TPR) and the false positive rate (FPR) in the test data for the following threshold probabilities: 1, 0.75, .5, 0.25, and 0?

2. You are given a dataset with 100 observations that has a continuous outcome y, and a single feature, x. You fit two OLS models. The first is a simple linear model i.e. regression of y on x. The second is a more flexible model in which you regress y on a cubic polynomial of x (i.e. x, x^2, x^3).
   a. Suppose the true relationship between y and x is linear ($y = \alpha + \beta x$). Would you expect the training MSE of your linear model to be lower than that of the cubic model? Explain your answer briefly.

   b. What would be your answer to part (a) if you were comparing the test MSE? Explain briefly.

   c. Suppose now that the true relationship between y and x is not linear, but we do not know the degree of non-linearity. Would you expect the training MSE of the linear model to be higher or lower than that of the cubic model? Explain briefly.

   d. What would be your answer to part (c) if you were comparing the test MSE? Explain briefly.

Short Answer True and False Questions

State whether the following statements are true or false, and explain your reason briefly (no more than 5 lines).

a. LASSO and Ridge regression coefficients can sometimes be identical to OLS regression coefficients.

b. Sometimes, one might pick a classification model with a Receiver Operating Characteristic (ROC) curve that lies below the 45-degree line.

c. Suppose you run the following constrained least squares regression (LASSO) with *p* predictors and *n* observations where we seek to minimize, for a particular value of *s*:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_p x_{ij})^2 \text{ subject to } \sum_{j=1}^{p}|\beta_j| \leq s$$

As we relax the constraint i.e. increase *s* from zero, the test MSE must decrease monotonically.

d. For the model in part c above, as we relax the constraint i.e. increase *s* from zero, the training MSE will decrease and then stay constant.