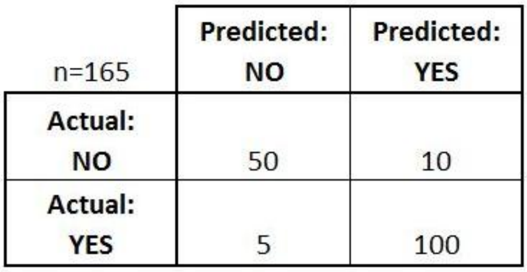
Quiz 2: Introduction to Data science: Tools and Techniques **Total Marks: 8** 2019-03-27

Name: --- Registration #: ----------------------- Section: ------

Q1. Imagine you are working on a project which is a binary classification problem. You trained a model on training dataset and get the below confusion matrix on validation dataset.

Based on the given confusion matrix, choose which option(s) below will give you correct predictions?

1. Accuracy is ~0.93
2. Misclassification rate is ~ 0.91
3. Precision is ~0.95
4. True positive rate (Recall) is ~0.95



n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

Reason (calculation):

Q2. For a specific choice of model, as the number of training points goes to infinity, describe the changes to the bias and variance exhibited by a model trained on the data.

Q3. Imagine, you are solving a classification problems with highly imbalanced class. The majority class is observed 99% of times in the training data. Your model has 99% accuracy after taking the predictions on test data. Which of the followings are true in such a case? (check all that apply)

A) Accuracy metric is not a good idea for imbalanced class problems.
B) Accuracy metric is a good idea for imbalanced class problems.
C) Precision and recall metrics are good for imbalanced class problems.
D) Precision and recall metrics aren't good for imbalanced class problems.

Q4. Suppose you train a regularized logistic regression classifier and test it. Training error is 500 (which is High), while test error is 510. Which of the following you should try to improve the performance of the algorithm. (check all that apply)

a) Try increasing Regularization parameter
b) Try decreasing Regularization parameter
c) Try adding new features
d) Try adding polynomial features
e) Try smaller number of features

Q5. You train a classifier and predict y=1 if h(x)>=0.5. On testing you came to know that the results (precision = 0.6, recall 0.65) are according to your requirements. Suppose you want to predict y=1 only if very confident, which of the followings are true:

a) You need to set the threshold higher (e.g., let say predict y=1 if h(x) >= 0.9)
b) You need to set the threshold lower (e.g., let say predict y=1 if h(x) >= 0.3)

c) Your precision will be greater than 0.6.

d) Your recall will be less than 0.65.

Q6. [1.5 marks] Suppose you have the following data of players and you need to fill the missing values of Height using single Imputation.

Fit the model and write down the values of θ_0 and θ_1.

a) Height of player “2”:

Player id	Height (inches)	Weight (kgs)
1	78	90
2		**83**
3	58	70

Q7. [1.5 Marks] A friend of yours is faced with a regression problem with two possible inputs, X_1 and X_2. he/she considers a linear regression model:

$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$

The data set is given in the following table:

X_1	X_2	y
2	2	18.1
3	3	24
1	4	15
5	3	32.9
4	2	20.1
6	6	23
5	4	12

Training Data: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), (x^{(7)}, y^{(7)})\}$

Cross Validation Data: $\{(x^{(6)}, y^{(6)})\}$

Test Data: $\{(x^{(4)}, y^{(4)}), (x^{(5)}, y^{(5)})\}$

Assuming three λ (Regularization parameter), the model is fitted to a training data set using mean-squared-errors, resulting in the three trained models respectively:

($\lambda = 0.1$) $h(x) = 10 - 2X_1 + 1.5X_2$
($\lambda = 0.3$) $h(x) = 11.6 - 1.4X_1 + 1.7X_2$
($\lambda = 0.5$) $h(x) = 2 + 2X_1 + 3X_2$

Your friend is puzzled by these results and comes to you for advice.

a) What do you think, which λ will be the best?
b) How well does the model generalize?

Solution: