Homework 6

Make a copy of this document or answer in a separate document.
(First two problems are the problems omitted from last week's homework)

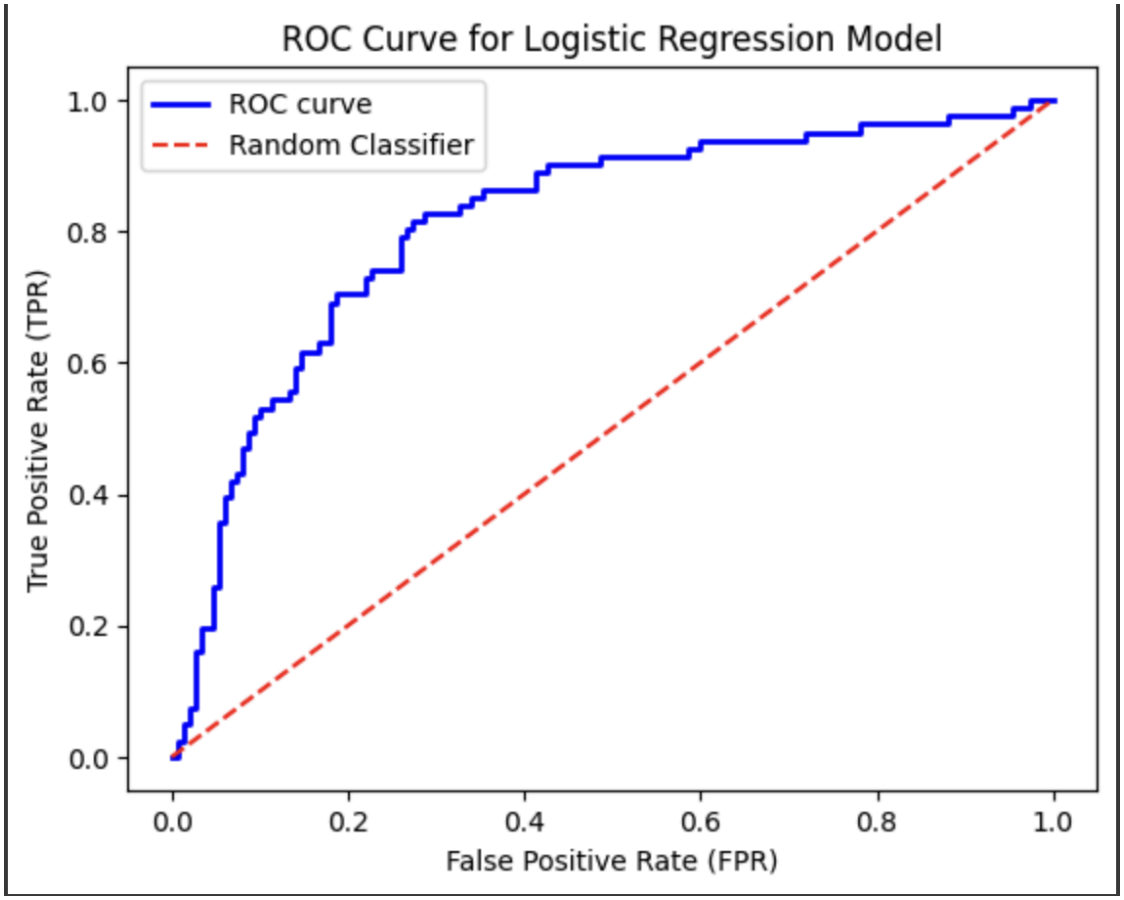
1. (1 point) Split the `pima_df` into testing and training and build a logistic regression model using the training set to predict "Outcome" being 1 given the other features. Summarize the model coefficients.

```
Optimization terminated successfully.
         Current function value: 0.452103
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                Outcome   No. Observations:                  537
Model:                          Logit   Df Residuals:                      528
Method:                           MLE   Df Model:                            8
Date:                Wed, 22 Oct 2025   Pseudo R-squ.:                 0.3005
Time:                        17:45:17   Log-Likelihood:                -242.78
converged:                       True   LL-Null:                       -347.09
Covariance Type:            nonrobust   LLR p-value:                9.750e-41
==============================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const                      -9.3830      0.924    -10.150      0.000     -11.195      -7.571
Pregnancies                 0.1303      0.040      3.227      0.001       0.051       0.209
Glucose                     0.0371      0.005      7.909      0.000       0.028       0.046
BloodPressure              -0.0137      0.007     -2.062      0.039      -0.027      -0.001
SkinThickness              -0.0072      0.009     -0.831      0.406      -0.024       0.010
Insulin                    -0.0010      0.001     -0.846      0.398      -0.003       0.001
BMI                         0.1026      0.019      5.423      0.000       0.066       0.140
DiabetesPedigreeFunction    1.3449      0.379      3.550      0.000       0.602       2.087
Age                         0.0226      0.012      1.881      0.060      -0.001       0.046
==============================================================================
```

2. (1 point) Using a 50% threshold, assign values with probability >= 0.5 a value of 1 and those lower to 0. Print the accuracy score and confusion matrix (confusion table) for these results.

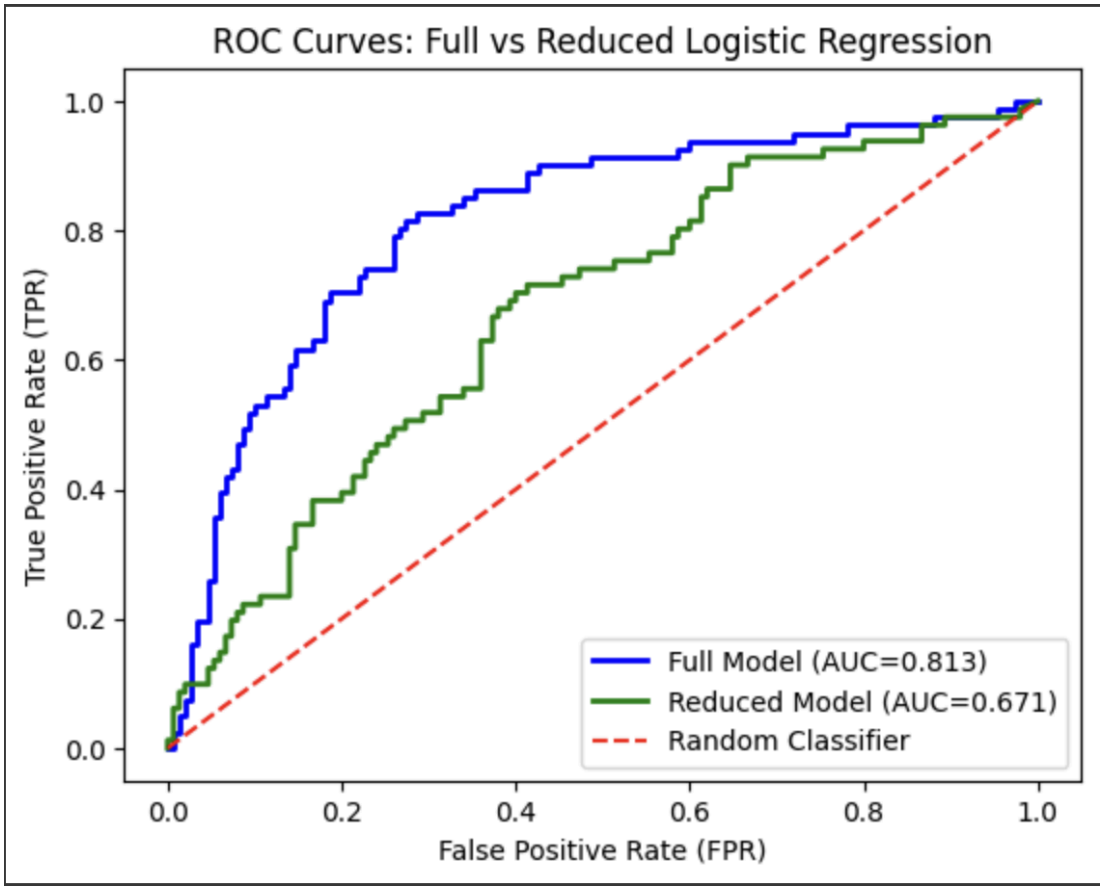
```
Accuracy: 0.766
Confusion Matrix:
Truth          0   1
Predicted
0            127  23
1             31  50
```

3. (1 point) Now we will try different thresholds by using sklearn's roc_curve function. Create an ROC curve for this model.



4. (2 points) Now create a new model which uses only "BMI", "BloodPressure", and "Insulin" to predict "Outcome". Create another ROC curve for this model. What are the AUC values for this model and the last model? Which is better?

```
Optimization terminated successfully.
         Current function value: 0.587415
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                Outcome   No. Observations:                  537
Model:                          Logit   Df Residuals:                      533
Method:                           MLE   Df Model:                            3
Date:                Wed, 22 Oct 2025   Pseudo R-squ.:                 0.09117
Time:                        18:12:09   Log-Likelihood:                -315.44
converged:                       True   LL-Null:                       -347.09
Covariance Type:            nonrobust   LLR p-value:                 1.164e-13
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          -4.3457      0.613     -7.085      0.000      -5.548      -3.143
BMI             0.0962      0.015      6.250      0.000       0.066       0.126
BloodPressure   0.0061      0.006      1.041      0.298      -0.005       0.018
Insulin         0.0015      0.001      1.931      0.053   -2.28e-05       0.003
==============================================================================
Full model AUC: 0.813
Reduced model AUC: 0.671
The full model performs better.
```



5. (1 point) If you had a very large amount of data would you use k-fold validation or LOOCV ? Why?
   I would use k-fold validation because it is not as computationally expensive as LOOCV and can handle larger amounts of data than LOOCV. Also, k-fold provides similar accuracy as LOOCV.

6. (1 point) What are "resampling" techniques and what are they used for?
   Resampling methods are an indispensable tool in modern statistics that involve repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain additional information about the fitted model. For example, it can obtain the variability of a model or access the performance.

7. (1 point) Define the following terms and connect which ones have similar meaning : "Sensitivity", "Specificity", "Type I Error", "Type II Error", "FPR", and "TPR"

   Sensitivity: the percentage of true positives that are identified; TPR
   Specificity: the percentage of true negatives that are identified; FPR
   Type I Error: if we erroneously reject H0 when H0 is in fact true; FPR
   Type II Error: if we do not reject H0 when H0 is in fact false; complement of TPR
   FPR: probability the test is positive when the condition is absent; Type I Error
   TPR: probability the test is positive when the condition is present; sensitivity

8. (1 point) Describe what is meant by *subset selection* and describe the different methods available.

   Subset selection is an approach that involves identifying a subset of the p predictors that we believe to be related to the response. Then the model is fit using least squares on the reduced set of variables. The different methods available include the "best subset selection" which involves fitting a separate least squares regressions for each possible combination of the p predictors and choosing the one that is the best. There are also stepwise selections which include forward stepwise, backward stepwise, and hybrid approaches.

9. (1 point) What is λ in shrinkage models? What values of λ have high variance and what values have high bias?
   λ in shrinkage models is also known as a tuning parameter and serves to control the relative impact of the two terms on the regression coefficient estimates. Smaller values have high variance and low bias while larger values have low variance and high bias.

Many problems should be solved using Python. The following notebook has been initialized with the packages and data that you need : link here

To turn in your assignment please take SCREENSHOTS of your results and insert them into (this or another) document for submission. You will then CONVERT THAT DOC TO A .PDF before submitting that. Please embed a hyperlink* to your google notebook INSIDE .pdf that you submit.

* : by "embed the link" I mean that I do NOT want you to copy/paste the disgusting hyperlink. Instead type a word, highlight the word and right click "insert link." Make sure it works! Then save as .pdf and the link should preserve!
Tran_HW6_Colab