

When you use inherently random methods, make sure to use **random_state=2023**. Please standardize your input before modeling if the model is not-scale invariant!

1. Consider the "Data1" data set. Let y be the response variable and $V_1, V_2, V_3, V_4 \dots V_8$ be the predictors. Use logistic regression in the following questions.
 - (a) Use **train_test_split** to randomly allocate 20% of the observations to the test set and the remaining 80% to the training set. Compute the cross-correlation between variables and display the correlation matrix using a **heat map**. Discuss your observations. (5%)
 - (b) Employ the training set to identify the best logistic regression model using three different methods: best subset selection, forward stepwise selection, and backward stepwise selection. In the first stage, when comparing models with an equal number of parameters, use [deviance](#) as the criterion. In the second stage, when comparing models with varying numbers of parameters, utilize **5-fold cross-validation** with [ROC AUC scoring](#). Which variables are selected by each of these three methods? Do they yield the same selections? (20%)
 - (c) Train a lasso logistic regression model on the training set and plot the coefficient estimates for the eight predictors against λ . Additionally: (7%)
 - Identify value of λ that minimizes the **5-fold cross-validation** error. When conducting cross-validation and plotting, scan λ (C) values ranging from 5×10^{-4} to 5×10^2 , using 200 evenly spaced samples on a logarithmic scale.
 - (d) Train a ridge logistic regression model on the training set and plot the coefficient estimates for the eight predictors against λ . Additionally: (7%)
 - Identify value of λ that minimizes the **5-fold cross-validation** error. When conducting cross-validation and plotting, scan λ (C) values ranging from 5×10^{-4} to 5×10^2 , using 200 evenly spaced samples on a logarithmic scale.
 - (e) Write down the final fitted model for the model selected by best subset selection and the lasso logistic regression model. Use the training set and test set to compute the training **ROC AUC** and test **ROC AUC** for these models. Does lasso logistic regression identify the same important variables as the best subset approach? (6%)
2. Consider the "Data2" data set. Let y be the response variable and $V_1, V_2, \dots V_{20}$ be the predictors. Please answer the following questions and use [weighted f1-score](#) to measure the performance.
 - (a) Use **train_test_split** to randomly allocate 20% of the observations to the test set, with the remaining 80% designated as the training set. Next, apply LDA for

- supervised dimensionality reduction, reducing the dimensionality to 2. Visualize the data points in this two-dimensional space using a scatter plot, distinguishing different classes with varied colors. What observations can you draw? (10%)
- (b) Use the training set to build the LDA classifier function. Find the training and test **weighted f1-score**. (5%)
 - (c) Use the training set to build the QDA classifier function. Find the training and test **weighted f1-score**. (5%)
 - (d) Use the training set to build the naive Bayes classifier function (Assuming each predictor is drawn from Gaussian distribution). Find the training and test **weighted f1-score**. (5%)
 - (e) Build the KNN classifier based on the training data set. Perform 10-fold cross validation to select the number of K (Please scan K from 1 to 20 with step size 1). Which value of K attains the largest **weighted f1-score**? Find the training and test **weighted f1-score** for the classifier with the selected number of K . (8%)
 - (f) Based on (b)-(e), which model will you suggest for prediction? Why? (5%)
3. Consider the "Data3" data set. Each row in the data set contains the related information of a real estate. Let **ePrice** be the response variable and $V_1, V_2, \dots, V_9$ be the predictors. Use linear regression to deal with the problem. Please answer the following questions. Be sure to handle the factor variable V_3 in your model.
- (a) Generate a **pair plot** for the dataset. Discuss your observations from this visualization. (5%)
 - (b) Utilize all the predictors to construct a multiple linear regression model. Is at least one of the predictors associated with the target **ePrice**? Justify your answer. Finally, explain the meaning of coefficients obtained in the model. (9%)
 - (c) Aim to enhance the model fit. Plot the leverage against studentized residuals. Additionally, using **Cook's distance**, identify the influential points (utilize the [OLSInfluence](#) module and consider $4/n$ as the threshold). Compare the fitting of the original model with those of the refitted model that excludes these influential points. What observations can you make? (9%)
 - (d) Augment the model from part (c) by adding second-degree (quadratic) polynomials of V_2 and V_5 . Does this addition enhance the model fit? (6%)
 - (e) Using the p -value and the hierarchy principle (chapter 3), which variables would you consider removing from the model in part (d)? Eliminate these variables and refit the model. What are the pros and cons when dropping these variables? (8%)