USC Viterbi
School of Engineering
Daniel J. Epstein Department of Industrial and Systems Engineering

ISE529 Predictive Analytics
Instructor: Dr. Tao Ma
2025 Spring

Homework 4
Due by: Apr. 6, 2025, 11:59 PM

Instructions:

1. Print your First and Last name and NetID on your answer sheets
2. Submit all your answers including Python scripts and report in a single Jupyter Lab file (.ipynb) or along with a single PDF to Brightspace by due date. No other file formats will be graded. No late submission will be accepted.
3. Total 4 problems. Total points: 100

---

1. (25 points)
Consider the use of a logistic regression model to predict the probability of *default* using *income* and *balance* on the Default data set. Compute estimates for the standard errors of the *income* and *balance* logistic regression coefficients in two different ways: (1) using the bootstrap method, and (2) using the standard function sm.GLM() or sm.Logit() from statsmodels library. Set a random seed = 0 when generate random indices for bootstrap.

(a) Using sm.GLM() or sm.Logit() function, determine the estimated standard errors for the coefficients associated with income and balance in a multiple logistic regression model

(b) Write a function, boot_fn(), that takes as input the *Default* data set as well as an index of the observations, and that outputs the coefficient estimates for income and balance in the multiple logistic regression model.

(c) Use your boot_fn() function to bootstrap 1000 samples to estimate the standard errors of the logistic regression coefficients for income and balance.

(d) Comment on the estimated standard errors obtained using the bootstrap and using sm.GLM() or sm.Logit() function.

2. (25 points)

Compute the LOOCV test error estimate for a simple logistic regression model on the Weekly data set. Write a "for" loop from $i = 1$ to n, where n is the number of observations in the data set, that performs each of the following steps:

i. Fit a logistic regression model with sm.GLM() function using all but the *i*th observation to predict *Direction* using *Lag1* and *Lag2*.

ii. Compute the probability of the market moving up with predict() function for the *i*th observation.

iii. Use the probability for the *i*th observation to predict whether the market moves up. *Pr*(*Direction* = "Up" | *Lag1, Lag2*) > 0.5.

iv. Determine the LOOCV test error estimate with the formula

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Err}_i,$$

where $Err_i = I\left(y_i \neq \hat{y}_i\right)$.

3. (25 points)
Consider the Carseats data set. The response *Sales* is a quantitative variable. Use random forests to analyze this data. Bootstrap 500 samples with random seed = 1. What training and test MSE do you obtain? Use the "feature_importances_" values to determine which variables are most important.

4. (25 points)
Use the Caravan data set to perform the following tasks:

(a) Create a training set consisting of the first 1,000 observations, and a test set consisting of the remaining observations.

(b) Fit a boosting model to the training set with *Purchase* as the response and the other variables as predictors. Use 1,000 trees, and a learning rate of 0.01, max splits of 4. Which predictors appear to be the most important? Show the list of importance.

(c) Use the boosting model to predict the response on the test data. Predict that a person will make a purchase if the estimated probability of purchase is greater than 20 %. Show a confusion matrix.