**DSI -06 Homework 4:**

Introduction with Statistical Learning with Applications to Python (ISLP) pg. 225-226

6. In Chapter 4, we used logistic regression to predict the probability of default using income and balance on the Default data set. We will now estimate the test error of this logistic regression model using the validation set approach. Do not forget to set a random seed before beginning your analysis.

1. Fit a logistic regression model that uses income and balance to predict default.
2. Using the validation set approach, estimate the test error of this model. In order to do this, you must perform the following steps:
   1. Split the sample set into a training set and a validation set.
   2. Fit a multiple logistic regression model using only the training observations.
   3. Obtain a prediction of default status for each individual in the validation set by computing the posterior probability of default for that individual, and classifying the individual to the default category if the posterior probability is greater than 0.5.
   4. Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.
3. Repeat the process in (b) three times, using three different splits of the observations into a training set and a validation set. Comment on the results obtained.
4. Now consider a logistic regression model that predicts the probability of default using income, balance, and a dummy variable for student. Estimate the test error for this model using the validation set approach. Comment on whether or not including a dummy variable for student leads to a reduction in the test error rate.

Additional Practice Questions:

Explain the concept of a random seed to a non-technical audience.

How would you describe this exercise in an interview to both a technical and non-technical interviewer? What are the key insights you would want to show?

Can you think of a business context where this exercise would have applications?