MIS 776
Assignment 4 – Classification

Data Scenario

Banks can generate significant profits from term deposits such as a certificate of deposit (CD). These deposits are required to be held for a certain period of time, which gives the bank access to those funds for lending purposes at a higher rate than the rate paid for the deposit. Of course, marketing term deposit products to customers can be expensive, so the bank will want to focus their efforts on those customers most likely to buy these products.

In this data set, we have information about 4521 customers, including demographic information as well as data related to their prior experience with the bank and previous marketing campaigns. Additionally, we have a class variable called *purchase* that indicates whether this customer purchased a term product in the current marketing campaign. Our objective is to predict which customers will purchase a term product if we spend the money to advertise to them. We want to develop a model that will maximize the returns based on the costs of marketing and the benefits of customer purchase. This data was adapted from a set published by Moro et al. (S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. *Decision Support Systems*, Elsevier, 62:22-31, June 2014)

Data

The target classification (output) column is *purchase*. All other columns are potential predictors.

Column Name	Definition
age	The age of the customer.
job	Job category of the customer.
marital	The marital status of the customer.
education	General education level (primary, secondary, tertiary).
default	Does the customer have a credit account in default?
balance	Current aggregate loan balance.
housing	Does the customer have a housing loan?
loan	Does the customer have a personal loan?
contact	Primary mode of contact with the customer.
day	The day of the month of last contact.
month	The month of last contact.
campaign	Number of contacts with customer during this campaign.
pdays	Number of days that passed after the client was last contacted from a previous campaign. (Note that this value is -1 if the customer was not contacted in a previous campaign)
previous	Number of total contacts before this campaign began.
poutcome	The outcome of previous marketing campaign.
purchase	Did the customer purchase a term deposit in this campaign?

Tasks (Justify your answers by providing screen shots or values you have observed that led to your conclusions.)

Submit answers to the following questions as well as the Python script that you used to get the answers.

1) Using the BankSet.csv file, open the file in a text editor (or a spreadsheet) and look at it. Make sure that you are comfortable with the contents of the file and the meanings of each column. Close the text editor when you are finished.
2) Using Python, read the contents of the file into a variable called *datBank*. What is the mean loan balance for the customers in our data set? What is the average number of contacts that have been made to a customer in this campaign?
3) Create a histogram of the loan balance (balance) and provide a screen shot of this histogram in your answer file. Is the distribution normal? Is it skewed?
4) Repeat the above for customer age.
5) Add a new column to your data set called *purchase_code*, which is a numeric value where 0 indicates that the customer did not purchase in the current campaign and 1 if they did purchase. As a check, the mean value of the new *purchase_code* column should be 0.1152, indicating that 11.52% of the customers **did** purchase the term deposit product.
6) Create a correlation table for all of the numeric values in the data set (age, balance, campaign, pdays, previous, purchase_code). Which numeric value is the most correlated with purchase_code? Which is the least correlated? Why is this important (why do we care about this)?
7) Generate aggregates of *purchase_code* for the following categorical predictors: job, marital, education, housing, loan, poutcome. Do any of these categorical predictors seem to have a high explanatory value for *purchase_code*? If so, which ones and why do you feel that this might be important?
8) Pick **one** of the categorical predictors that you think should have the greatest impact on purchase. Use one-hot encoding to transform this categorical variable into numeric values for analysis.
9) Store the columns that are numeric and the ones that you one-hot encoded into a data set called 'X'. This will be the feature data set. This should contain only numeric values and it should not contain purchase_code.
10) Store the purchase_code values in a data set called 'y'. This will be the target data set that you are trying to predict.
11) Randomly partition the rows into training and validation set using a random seed of 500. Put 70% into the training partition and 30% into the testing partition.
12) Apply a decision tree classifier to the training data to create a model. Calculate the Accuracy % of the model as a whole. Create a confusion matrix and calculate the Precision, Recall, and F-Measure for each purchase class (yes and no). Since the distribution of yes to no in this set is 11.5%, how well do you feel that this model does at predicting campaign purchasers? What is the primary predictor? Do you think that using this predictor exclusively might be potentially problematic?
13) Use the validation (testing) partition to run predictions on the model and generate a confusion matrix on the test set. Is there any indication of overfit or any other problems that you see?
14) Repeat steps 12 and 13 above using a naïve Bayes classifier.
15) Which of the two models (tree or Bayesian) do you think is a better approach for this data set and why?