# FIN 70234/40230 Prof. Barry Keating

# Business Forecasting

#### K-Nearest Neighbor Data Mining Exercise #2

**Purpose:** To learn how to choose a “good” **K-Nearest Neighbors classification model** by **minimizing, over a reasonable number of neighborhood sizes (k) and probability cutoff values (p(cutoff)), the total misclassification error percentage based on the validation data set**.

Go to the website for this course and download the file “**Gatlin2data.xls**”. Use it in conjunction with XLMiner © to answer the following questions. Hand in your work on the required date.

We are going to build a K-Nearest Neighbors classification model for the Gatlin data. The classification variable y takes the value of 1 if the tested real estate agent subsequently became “successful” and zero otherwise. For the definitions of the explanatory variables proposed, see the description provided in the Gatlin2data.xls file. Partition all of the Gatlin data into two parts: training (60%) and validation (40%). We won’t use a test data set this time. Use the default random number seed 12345.

Using this partition, we are going to build a K-Nearest Neighbors classification model using **all** (8) of the available input variables (i.e., the “X” variables). For the K-Nearest Neighbors classification model there are **two tuning parameters**: k, the number of neighbors, and p (cutoff), the probability value used to determine if a candidate in the validation dataset is to be judged a “success” or a “failure.”

If the K-Nearest Neighbors classification model predicts the probability of success of a new case (observation) to be greater or equal to the pre-specified probability cutoff value, p (cutoff), then the new case is predicted to be a success (=1). Otherwise, the new case is predicted to be a failure (= 0). Usually p (cutoff) is set equal to 0.5 but, in some instances, a p (cutoff) value slightly higher than 0.5 (say, 0.6) or a p (cutoff) value slightly less than 0.5 (say, 0.4) might provide a better set of classification predictions on the validation dataset than just using the standard p (cutoff) = 0.5 value.

We proceed to build a K-Nearest Neighbors classification model in the following sequential manner. In the validation dataset, we first are going to tune on the number of nearest neighbors (k), while holding p (cutoff) = 0.5 by minimizing the total misclassification percentage.

Then, once we find a “good” number of nearest neighbors, say k\*, we are going to hold k = k\* and tune over the p(cutoff) until we a “best” K-Nearest neighbors classification model that minimizes the total misclassification percentage over the validation dataset for our reasonable choices of k and p(cutoff). (**Remember to normalize your data when building your models**.)

1. **Using the validation data set classification scores**, fill in the following table:

### # of Nearest Neighbors p (cutoff) Total % Misclassification Error

3 0.5

5 0.5

7 0.5

9 0.5

12 0.5

15 0.5

Given p (cutoff) = 0.5, what is the best number of nearest neighbors k\* = \_\_\_?

Explain your answer.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

1. Using the best k = k\* determined in part a) above, fill in the following table, **using the validation data set classification scores**.

### # of Nearest Neighbors p (cutoff) Total % Misclassification Error

k\* = \_\_\_\_ 0.4

k\* = \_\_\_\_ 0.6

What is the best tuning value for p (cutoff)? 0.5, 0.4, or 0.6?

What is the best Total % Misclassification Error = \_\_\_\_\_\_\_?

What is the best Nearest-Neighbors classification model for the Gatlin dataset? k\* = \_\_\_,

p (cutoff)\* = \_\_\_\_.

Recall that the estimated ***misclassification rate*** (also called the ***overall*** ***error rate***) is given by the total misclassifications divided by the total number of cases.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

1. For the very best model determined above, print out the validation data set **“traditional” Lift Chart** and the **“decile-wise” Lift Chart** and hand it in with this exercise. **Explain the interpretations of these two charts**. At what point in the “traditional” Lift Chart is the “lift” the maximum? Hint: Put your pointer on the maximum point and the point will be revealed on the screen.