**+++++++++**

**>> Problem #1:** A large number of insurance records are to be examined to develop a model for predicting fraudulent claims. Of the claims in the historical database, 1% were judged to be fraudulent.

A sample database is taken to develop a model, and oversampling is used to provide a balanced sample in light of the very low response rate. When applied to this sample database (total number of records, N =800), the model ends up correctly classifying 310 frauds, and 270 non-frauds. It misses 90 frauds, and classified 130 records incorrectly as frauds when they were not.

**Questions:**

**1a.)** Produce the classification matrix for the sample as it stands.

|  |  |  |
| --- | --- | --- |
| **As is** | **Fraud (Actual)** | **Non-Fraud (Actual)** |
| **Fraud** | 310 | 130 |
| **Non-Fraud** | 90 | 270 |

**1b.)** Find the adjusted misclassification rate (adjusting for the oversampling).

|  |  |  |
| --- | --- | --- |
| **Adjusted** | **Fraud (Actual)** | **Non-Fraud (Actual)** |
| **Fraud** | 6.2 | 257.4 |
| **Non-Fraud** | 1.8 | 534.6 |

= 0.32

**1c.)** What percentage of new records would you expected to be classified as fraudulent?

**+++++++++**

**>> Problem #2:** The file **BostonHousing.xls** **(**See **\*\*attached** **file)** contains information on over 500 census tracts in Boston, where for eachtract 14 variables are recorded. The last column (CAT.MEDV) was derived from MEDV, such that it obtains the value 1 if MEDV>30 and 0 otherwise. Consider the goal of predicting the median value (MEDV) of a tract, given the information in the first 13 columns. Partition the data into training (60%) and validation (40%) sets. **(**For interpretation of the column names in BostonHousing.xls,

reference the **\*\*attached Table 2.2, shown below)**

**Questions:**

**2a.)** Perform a k-nearest neighbors prediction with all 13 predictors (the CAT.MEDV column is the outcome or decision variable), trying values of k from 1 to 10. Make sure to normalize the data (click “normalize input data"). What is the best k chosen? What does it mean?

|  |  |
| --- | --- |
| **K** | **Total Sum of squared Error** |
| **1** | **16** |
| **2** | **12.14** |
| **3** | **10.16** |
| **4** | **10.51** |
| **5** | **9.96** |
| **6** | **9.62** |
| **7** | **9.19** |
| **8** | **8.92** |
| **9** | **9.19** |
| **10** | **9.52** |

**The 8th nearest neighbor seems to be the optimal number for k based on the cross-validation. It means that if we use more than 8 points to classify the data, we are missing some nuances. After 8, we are basically overfitting the model to data.**

**2b.)** Why is the validation data error overly optimistic compared to the error rate when applying this k-NN predictor to new data?

Because the classifier is trained on this dataset and it is overfitted to our data. Our partitioning while random is still from the same data base.

**+++++++++**

**>>** \*\*Table 2.2**:**

