Exercises

1. A data mining algorithm has been applied to a transaction dataset and has classified 88 records as fraudulent (30 correctly so) and 952 as nonfraudulent (920 correctly so). Which of the following situations represents the confusion matrix for the transactions data mentioned? Explain your reasoning.

A.

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	58	920
0	30	32

B.

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	32	30
0	58	920

C.

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	30	32
0	58	920

D.

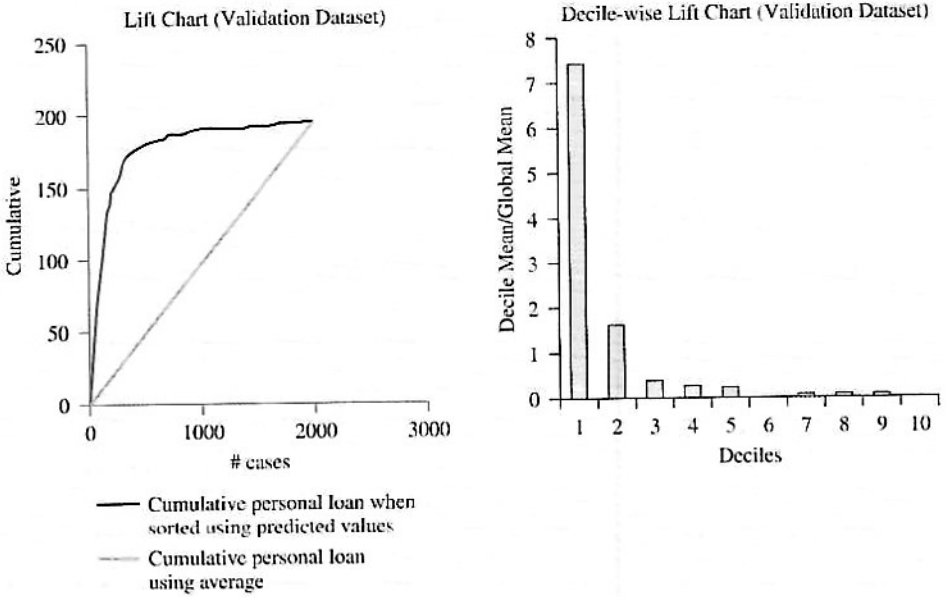
Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	920	58
0	30	32

2. Calculate the classification error rate for the following confusion matrix. Comment on the pattern of misclassifications. How much better did this data mining technique do compared to a naive model? What is the misclassification rate for the naive model?

	Predict Class 1	Predict Class 0
Actual 1	8	2
Actual 0	20	970

3. Explain what is meant by Bayes' theorem as used in the Naive Bayes model.

4. Explain the difference between a training data set and a validation data set. Why are these data sets used routinely with data mining techniques in the XLMiner® program and not used in the ForecastX™ program? Is there, in fact, a similar technique presented in a previous chapter that is much the same as partitioning a data set?
5. For a data mining classification technique, the validation data set lift charts are shown below. What confidence in the model would you express, given this evidence?



Source: Frontline Systems Inc.

6. Wine (in this case, red wine) has been graded for many years by experts who actually taste a sample of the wine, examine its color and aroma, and assign a grade (in our case, high quality or lower quality). Would it be possible, however, to use attributes of the wine that are machine measurable such as fixed acidity and residual sugar to classify the wines?

   Partition the data into two partitions (60 percent and 40 percent, respectively), the training data and the validation data. Estimate a kNN algorithm for the training data, and examine the resulting estimation.

   Explain the overall misclassification rate for the validation partition and its calculation. What would have been the validation misclassification rate if you had used the naive model?

   What does the validation confusion matrix tell the data scientist? Is the algorithm more likely to make one type of error than the other?

   The lift chart is potentially the most important information provided by the algorithm. Examine either the cumulative gains chart, the lift chart proper, or the decile-wise lift chart. Do they each tell much the same story? How would you explain one of these lift charts to someone unfamiliar with predictive analytics?

7. Use the red wine data again with the same partition and estimate a classification model with Naive Bayes.

   Again explain the overall misclassification rate for the validation data. Is it different for the Naive Bayes algorithm and the kNN algorithm?

   Is the confusion matrix for the validation data different than the one obtained with the kNN algorithm?

   Finally, examine the lift chart and make a comparison with the kNN model.

8. The bank marketing data includes actual information for a direct marketing effort by a bank. We will attempt to construct a model with just a few of the available attributes. We are interested in classifying whether a customer will respond positively to the marketing effort offering a term deposit.

   The attributes you are to use are age, balance, duration, campaign, pdays, and previous; explanations of these numeric variables are in the file. After your initial analysis, you may wish to transform some of the remaining variables to attempt to estimate a more complete model.

   Use a logit model for the estimate, making sure to request an "analysis of coefficients" in XLMiner®. As usual, use a 60/40 split for the training and validation data sets, and request a full set of lift charts.

   Does this estimate, using only some of the available attributes, do better than a naive model in the overall misclassification rate? What if you examine the lift chart? Recall that the lift chart reorders the data from most likely to accept a marketing offer to least likely to accept such an offer. Now does the algorithm appear to have explanatory power (i.e., could you successfully use it to suggest who to market to in the first place)?

   Which of the attributes that you selected appear to have the greatest effect on the classification? How certain are you that these attributes have an effect on the classification?

   By creating dummy variables and categorical variables for the attributes that you did not use in this exercise already, you may extend the analysis in order to refine the algorithm. Evaluate the resulting output in the same manner described above and compare the two outputs. Did the addition of the extra attributes to the logit model add additional explanatory power?

9. The Boston housing data includes information from the 1970 U.S. census for the city of Boston and surrounding area. Note that there are two variables representing value; one is "Medv," which is a dollar value. The other is "Cat Medv," which is a binary variable indicating whether the house is of "high value" (signified by "1") or "lower value" (identified by "0").

   Estimate a classification tree using the Cat Medv variable as the target. You are trying to classify home as either high value or lower value by using the given attributes (such as the number of rooms in the house, the crime rate in the local area, and the age of the houses in the area). Use all of the attributes in the file to estimate the model requesting the "best pruned tree" to prevent overfitting. For display, however, request the "full tree."

   Evaluate the estimate for the best pruned tree using the confusion matrix, the misclassification rate, and, most importantly, the lift chart.

   By examining the full tree, you should be able to see how the CART algorithm will attempt to perfectly classify the records. In doing so, it may overfit the data, and that is the reason for using the pruning method.

   Now re-estimate the algorithm using the target Medv this time. In order to do so, you will have to use the "Prediction" menu in XLMiner® and select "Regression Tree."

The method is similar to the classification tree estimated earlier, but now an actual numerical prediction is being requested. Overfitting remains a possible problem, and so it will again be necessary to prune using either the best pruned tree or the minimum error tree.

10. The credit card fraud data is a small version (comprised of 14,240 records) of a much larger data set (containing 248,807 records); it is made up of 2013 European transactions. It is a very unbalanced data set in which there are only a few fraudulent transactions. Attempting to classify transactions as fraudulent will be difficult since there are very few instances of fraud.

    Use logit and a kNN model to create a predictive model for the credit card fraud data. Does either of these models have predictive power?

    Explain carefully the information provided by the lift chart or the decile-wise lift chart; how does this information differ from the information provided by the overall misclassification rate?

    What value to a firm could you see in creating such a model and using it in real time?