Captured from Amazon Preview of the Book. This will be provided **ONLY** for the 1st homework assignment but will **NOT** be provided for future homework assignments. Students **MUST** have their own textbook for this class.

PROBLEMS

2.1 Assuming that data mining techniques are to be used in the following cases, identify whether the task required is supervised or unsupervised learning.

- **a.** Deciding whether to issue a loan to an applicant based on demographic and financial data (with reference to a database of similar data on prior customers).
- **b.** In an online bookstore, making recommendations to customers concerning additional items to buy based on the buying patterns in prior transactions.
- **c.** Identifying a network data packet as dangerous (virus, hacker attack) based on comparison to other packets whose threat status is known.
- **d.** Identifying segments of similar customers.
- **e.** Predicting whether a company will go bankrupt based on comparing its financial data to those of similar bankrupt and nonbankrupt firms.
- **f.** Estimating the repair time required for an aircraft based on a trouble ticket.
- **g.** Automated sorting of mail by zip code scanning.
- **h.** Printing of custom discount coupons at the conclusion of a grocery store checkout based on what you just bought and what others have bought previously.

2.2 Describe the difference in roles assumed by the validation partition and the test partition.

2.3 Consider the sample from a database of credit applicants in Table 2.15. Comment on the likelihood that it was sampled randomly, and whether it is likely to be a useful sample.

TABLE 2.15 SAMPLE FROM A DATABASE OF CREDIT APPLICATIONS

OBS	CHECK ACCT	DURATION	HISTORY	NEW CAR	USED CAR	FURNITURE	RADIO TV	EDUC	RETRAIN	AMOUNT	SAVE ACCT	RESPONSE
1	0	6	4	0	0	0	1	0	0	1169	4	1
8	1	36	2	0	1	0	0	0	0	6948	0	1
16	0	24	2	0	0	0	1	0	0	1282	1	0
24	1	12	4	0	1	0	0	0	0	1804	1	1
32	0	24	2	0	0	1	0	0	0	4020	0	1
40	1	9	2	0	0	0	1	0	0	458	0	1
48	0	6	2	0	1	0	0	0	0	1352	2	1
56	3	6	1	1	0	0	0	0	0	783	4	1
64	1	48	0	0	0	0	0	0	1	14421	0	0
72	3	7	4	0	0	0	1	0	0	730	4	1
80	1	30	2	0	0	1	0	0	0	3832	0	1
88	1	36	2	0	0	0	0	1	0	12612	1	0
96	1	54	0	0	0	0	0	0	1	15945	0	0
104	1	9	4	0	0	1	0	0	0	1919	0	1
112	2	15	2	0	0	0	0	1	0	392	0	1

2.4 Consider the sample from a bank database shown in Table 2.16; it was selected randomly from a larger database to be the training set. *Personal Loan* indicates whether a solicitation for a personal loan was accepted and is the response variable. A campaign is planned for a similar solicitation in the future and the bank is looking for a model that will identify likely responders. Examine the data carefully and indicate what your next step would be.

Copyrighted Material

TABLE 2.16 SAMPLE FROM A BANK DATABASE

OBS	AGE	EXPERIENCE	INCOME	ZIP CODE	FAMILY	CC AVG	EDUC	MORTGAGE	PERSONAL LOAN	SECURITIES ACCT
1	25	1	49	91107	4	1.6	1	0	0	1
4	35	9	100	94112	1	2.7	2	0	0	0
5	35	8	45	91330	4	1	2	0	0	0
9	35	10	81	90089	3	0.6	2	104	0	0
10	34	9	180	93023	1	8.9	3	0	1	0
12	29	5	45	90277	3	0.1	2	0	0	0
17	38	14	130	95010	4	4.7	3	134	1	0
18	42	18	81	94305	4	2.4	1	0	0	0
21	56	31	25	94015	4	0.9	2	111	0	0
26	43	19	29	94305	3	0.5	1	97	0	0
29	56	30	48	94539	1	2.2	3	0	0	0
30	38	13	119	94104	1	3.3	2	0	1	0
35	31	5	50	94035	4	1.8	3	0	0	0
36	48	24	81	92647	3	0.7	1	0	0	0
37	59	35	121	94720	1	2.9	1	0	0	0
38	51	25	71	95814	1	1.4	3	198	0	0
39	42	18	141	94114	3	5	3	0	1	1
41	57	32	84	92672	3	1.6	3	0	0	1