# **2.2**

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

When looking at the different roles, the validation partition is used to understand and evaluate the performance of the prediction model. This allows you to compare models and choose the best one. The test partition is used to compare against the validation partition and often is used in automated ways to tune and improve the model.

# **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.

The sample shows that all of the OBS values are multiples of 8 which indicate that it is not a random sample and likely is derived using a math equation, therefore, most likely not to be a useful sample.

# **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.

My observation shows that the chance of loan approval odds for someone in the age group less than 45 to be a higher chance that they will receive a loan was accepted. In order to accurately decide whether this is true more data and prediction models would be needed.

# **2.5**

Using the concept of overfitting, explain why when a model is fit to training data, zero error with those data is not necessarily good.

The book does a great job in explaining the answer to this question, “As a result, the general approach to data mining is vulnerable to the danger of overfitting, where a model is fit so closely to the available sample of data that it describes not merely structural characteristics of the data, but random peculiarities as well. In engineering terms, the model is fitting the noise, not just the signal.” Data Mining for Business Analytics Concepts, Techniques, and Applications in R by Galit Shmueli

In other words, the data may be “noisy” and may not always be a good thing.

# **2.6**

In fitting a model to classify prospects as purchasers or nonpurchases, a certain company drew the training data from internal data that include demographic and purchase information. Future data to be classified will be lists purchased from other sources, with demographic (but not purchase) data included. It was found that “refund issued” was a useful predictor in the training data. Why is this not an appropriate variable to include in the model?

Because the fitting model was created using two categories when it should have been created using three, this will have a negative approach to the model. What should have happened is a third category should have been created and them grouped together to show what products are going to be returned.

# **2.7**

A dataset has 1000 records and 50 variables with 5% of the values missing, spread randomly throughout the records and variables. An analyst decides to remove records with missing values. About how many records would you expect to be removed?

1000 records x 50 variables = 50000 attributes. .05% 🡪 missing = 2500 missing. 2500/50colums = 50 records removed.

# **2.8**

**A screenshot of a computer

Description automatically generated with medium confidence**

# **2.9**

Statistical distance between records can be measured in several ways. Consider Euclidean distance, measured as the square root of the sum of the squared differences. For the first two records in Table 2.17, it is √ (25 − 56)2 + (49, 000 − 156, 000)2. PROBLEMS 51 Can normalizing the data change which two records are farthest from each other in terms of Euclidean distance?

Normalizing the data would change which records are furthest from each other. Because the Euclidean distance is the length of a line segment between two points, normalizing the data would cause this distance to change and impact the measurement of the data.

# **2.10**

Two models are applied to a dataset that has been partitioned. Model A is considerably more accurate than model B on the training data, but slightly less accurate than model B on the validation data. Which model are you more likely to consider for final deployment?

Model B looks like the better choice because the data used in Model A is less accurate.