Questions

What are the key differences between the data mining process adopted by the article’s authors (shown in Figure 1 in the article) with the CRISP-DM process described in the textbook?

In their discussion of the data mining process, what key points emphasized by the article’s authors echo key points about the process made by the authors of DSFB?

What type of data mining were the article’s authors doing, supervised or unsupervised? Explain your answer.

What do you think the article’s authors mean when they state on page 273 that companies should not embark on data mining “projects”?

Why does the data preparation phase typically require a large amount of time and resources? (Make sure to use some specifics from the article in your response. Keep in mind that the authors divide the CRISP-DM data preparation phase into data warehouse and data preparation).

What is the difference between inferential statistics and data mining?

What was the target variable for the data mining effort?

Could this data mining effort be referred to as “Big Data Analytics”? Why or why not?

For the modeling task they retained only records that appeared in all three files (household file, promotion file, transaction file). Why does this have potential to undermine the conclusions of their modeling effort? (Hint: This is a challenging question, and the answer is not in the article. To answer it, think about what you learned in statistics about populations and representative samples.)

Why did they choose to do the modeling with a decision tree instead of regression or a neural network?

Why was the data randomly split into three datasets?

On page 277 you will find the sentence “The chi-square statistic was used to evaluate candidate splits with a maximum acceptable p value for each split of 0.2.” What does this mean? (Hint: You will probably need to Google this one, with combination of search terms like “decision tree,” “candidate splits,” and “chi-square.”)

After considering 100 variables only 22 were used as input variables for the decision tree. Why do you think they limited it to 22 variables? How do you think they decided which 22 variables to use? (Hint: These are thought questions. The answers are not in the article.)

AutoPay is one of the variables on which the decision tree splits. We don’t know all the details of the data, but can you see a potential problem with using the AutoPay variable from this historical data to predict likelihood of churn?

Why do the article authors suggest that a promotion should be tried on customers reaching tenure of around 30 months?