APPENDIX A
Proposal Review Guide

Effective data analytic thinking should allow you to assess potential data mining projects systematically. The material in this book should give you the necessary background to assess proposed data mining projects, and to uncover potential flaws in proposals. This skill can be applied both as a self-assessment for your own proposals and as an aid in evaluating proposals from internal data science teams or external consultants.

What follows contains a set of questions that one should have in mind when considering a data mining project. The questions are framed by the data mining process discussed in detail in Chapter 2, and used as a conceptual framework throughout the book. After reading this book, you should be able to apply these conceptually to a new business problem. The list that follows is not meant to be exhaustive (in general, the book isn't meant to be exhaustive). However, the list contains a selection of some of the most important questions to ask.

Throughout the book we have concentrated on data science projects where the focus is to mine some regularities, patterns, or models from the data. The proposal review guide reflects this. There may be data science projects in an organization where these regularities are not so explicitly defined. For example, many data visualization projects initially do not have crisply defined objectives for modeling. Nevertheless, the data mining process can help to structure data-analytic thinking about such projects—they simply resemble unsupervised data mining more than supervised data mining.

Business and Data Understanding

- What exactly is the business problem to be solved?
- Is the data science solution formulated appropriately to solve this business problem? *NB: sometimes we have to make judicious approximations.*
- What business entity does an instance/example correspond to?

- Is the problem a supervised or unsupervised problem?
  - If supervised,
    - Is a *target* variable defined?
    - If so, is it defined precisely?
    - Think about the values it can take.
- Are the attributes defined precisely?
  - Think about the values they can take.
- For supervised problems: will modeling this target variable improve the stated business problem? An important subproblem? If the latter, is the rest of the business problem addressed?
- Does framing the problem in terms of expected value help to structure the subtasks that need to be solved?
- If unsupervised, is there an "exploratory data analysis" path well defined? (That is, *where is the analysis going?*)

Data Preparation

- Will it be practical to get values for attributes and create feature vectors, and put them into a single table?
- If not, is an alternative data format defined clearly and precisely? Is this taken into account in the later stages of the project? (Many of the later methods/techniques assume the dataset is in feature vector format.)
- If the modeling will be supervised, is the target variable well defined? Is it clear how to get values for the target variable (for training and testing) and put them into the table?
- How exactly will the values for the target variable be acquired? Are there any costs involved? If so, are the costs taken into account in the proposal?
- Are the data being drawn from the similar population to which the model will be applied? If there are discrepancies, are the selection biases noted clearly? Is there a plan for how to compensate for them?

Modeling

- Is the choice of model appropriate for the choice of target variable?
  - Classification, class probability estimation, ranking, regression, clustering, etc.

- Does the model/modeling technique meet the other requirements of the task?
  - Generalization performance, comprehensibility, speed of learning, speed of application, amount of data required, type of data, missing values?
  - Is the choice of modeling technique compatible with prior knowledge of problem (e.g., is a linear model being proposed for a definitely nonlinear problem)?
- Should various models be tried and compared (in evaluation)?
- For clustering, is there a similarity metric defined? Does it make sense for the business problem?

Evaluation and Deployment

- Is there a plan for domain-knowledge validation?
  - Will domain experts or stakeholders want to vet the model before deployment? If so, will the model be in a form they can understand?
- Is the evaluation setup and metric appropriate for the business task? Recall the original formulation.
  - Are business costs and benefits taken into account?
  - For classification, how is a classification threshold chosen?
  - Are probability estimates used directly?
  - Is ranking more appropriate (e.g., for a fixed budget)?
  - For regression, how will you evaluate the quality of numeric predictions? Why is this the right way in the context of the problem?
- Does the evaluation use holdout data?
  - Cross-validation is one technique.
- Against what baselines will the results be compared?
  - Why do these make sense in the context of the actual problem to be solved?
  - Is there a plan to evaluate the baseline methods objectively as well?
- For clustering, how will the clustering be understood?
- Will deployment as planned actually (best) address the stated business problem?
- If the project expense has to be justified to stakeholders, what is the plan to measure the final (deployed) business impact?