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?

APPENDIX B

Another Sample Proposal

Appendix A presented a set of guidelines and questions useful for evaluating data science proposals. Chapter 13 contained a sample proposal ("Example Data Mining Proposal" on page 325) for a "customer migration" campaign and a critique of its weaknesses ("Flaws in the Big Red Proposal" on page 326).

We've used the telecommunications churn problem as a running example throughout this book. Here we present a second sample proposal and critique, this one based on the churn problem.

Scenario and Proposal

You've landed a great job with Green Giant Consulting (GGC), managing an analytical team that is just building up its data science skill set. GGC is proposing a data science project with TelCo, the nation's second-largest provider of wireless communication services, to help address their problem of customer churn. Your team of analysts has produced the following proposal, and you are reviewing it prior to presenting the proposed plan to TelCo. Do you find any flaws with the plan? Do you have any suggestions for how to improve it?

> **Churn Reduction via Targeted Incentives — A GGC Proposal**
>
> We propose that TelCo test its ability to control its customer churn via an analysis of churn prediction. The key idea is that TelCo can use data on customer behavior to predict when customers will leave, and then can target these customers with special incentives to remain with TelCo. We propose the following modeling problem, which can be carried out using data already in TelCo's possession.
>
> We will model the probability that a customer will (or will not) leave within 90 days of contract expiration, with the understanding that there is a separate problem of retaining customers who are continuing their service month-to-month, long after contract expiration. We believe that predicting churn in this 90-day window is an appropriate starting point, and the lessons learned may apply to other churn-prediction cases as well. The

> model will be built on a database of historical cases of customers who have left the company. Churn probability will be predicted based on data 45 days prior to contract expiration, in order for TelCo to have sufficient lead time to affect customer behavior with an incentive offer. We will model churn probability by building an ensemble of trees (random forest) model, which is known to have high accuracy for a wide variety of estimation problems.
>
> We estimate that we will be able to identify 70% of the customers who will leave within the 90-day time window. We will verify this by running the model on the database to verify that indeed the model can reach this level of accuracy. Through interactions with TelCo stakeholders, we understand that it is very important that the V.P. of Customer Retention sign off on any new customer retention procedures, and she has indicated that she will base her decision on her own assessment that the procedure used for identifying customers makes sense and on the opinions about the procedure from selected firm experts in customer retention. Therefore, we will give the V.P. and the experts access to the model, so that they can verify that it will operate effectively and appropriately. We propose that every week, the model be run to estimate the probabilities of churn of the customers whose contracts expire in 45 days (give or take a week). The customers will be ranked based on these probabilities, and the top *N* will be selected to receive the current incentive, with *N* based on the cost of the incentive and the weekly retention budget.

Flaws in the GGC Proposal

We can use our understanding of the fundamental principles and other basic concepts of data science to identify flaws in the proposal. Appendix A provides a starting "guide" for reviewing such proposals, with some of the main questions to ask. However, this book as a whole really can be seen as a proposal review guide. Here are some of the most egregious flaws in Green Giant's proposal:

1. The proposal currently only mentions modeling based on "customers who have left the company." For training (and testing) we will also want to have customers who did *not* leave the company, in order for the modeling to find discriminative information. (Chapter 2, Chapter 3, Chapter 4, Chapter 7)
2. Why rank customers by the highest probability of churn? Why not rank them by expected loss, using a standard expected value computation? (Chapter 7, Chapter 11)
3. Even better, should we not try to model those customers who are most likely to be influenced (positively) by the incentive? (Chapter 11, Chapter 12)
4. If we're going to proceed as in (3), we have the problem of not having the training data we need. We'll have to invest in obtaining training data. (Chapter 3, Chapter 11)

Note that the current proposal may well be just a first step toward the business goal, but this would need to be spelled out explicitly: *see if we can estimate the probabilities well.* If we can, then it makes sense to proceed. If not, we may need to rethink our investment in this project.

5. The proposal says nothing about assessing *generalization* performance (i.e., doing a holdout evaluation). It sounds like they are going to test on the training set ("…running the model on the database…"). (Chapter 5)
6. The proposal does not define (nor even mention) what attributes are going to be used! Is this just an omission? Is this because the team hasn't even thought about it? What is the plan? (Chapter 2, Chapter 3)
7. How does the team estimate that the model will be able to identify 70% of the customers who will leave? There is no mention that any pilot study already has been conducted, nor learning curves having been produced on data samples, nor any other support for this claim. It seems like a guess. (Chapter 2, Chapter 5, Chapter 7)
8. Furthermore, without discussing the error rate or the notion of false positives and false negatives, it's not clear what "identify 70% of the customers who will leave" really means. If I say nothing about the false-positive rate, I can identify 100% of them simply by saying everyone will leave. So talking about true-positive rate only makes sense if you also talk about false-positive rate. (Chapter 7, Chapter 8)
9. Why choose one particular model? With modern toolkits, we can easily compare various models on the same data. (Chapter 4, Chapter 7, Chapter 8)
10. The V.P. of Customer Retention must sign off on the procedure, and has indicated that she will examine the procedure to see if it makes sense (domain knowledge validation). However, ensembles of trees are black-box models. The proposal says nothing about how she is going to understand how the procedure is making its decisions. Given her desire, it would be better to sacrifice some accuracy to build a more comprehensible model. Once she is "on board" it may be possible to use less-comprehensible techniques to achieve higher accuracies. (Chapter 3, Chapter 7, Chapter 12)