Review 1, chs 1, 3-9

It is particularly hard to evaluate a book from only the few first chapters. From what I read, this book seems to address analysts who are already familiar with statistics and/or machine learning methods, but lack knowledge on graphical tools to evaluate their predictive models. I have a hard time to see it used as a textbook because I found the book a little bit too superficial to be used in a course. It is more a "recipe book" than a monograph, at the point that I had the impression, at least from the chapters I read, to have to do with an extended vignette of the R package "ingredients". I do not deny the advantages of such approach, especially for a book whose goal is to promote the use of specific visual tools for evaluating/understanding predictive models, but, as an academic, I probably cannot fully appreciate it.

Some points I noticed:

- the concept of prediction model, which is the key concept of the book, is not really introduced, but very vaguely in the first sentence of Chapter 1. Although the focus is on the graphical tools and their use, I think it is important to clearly introduce the basic concepts in a book. In addition, the authors completely ignore the separation between models for prediction and models for explanation. It may be fine, depending on the audience, but it is important to not mix up the concepts, at least without a real discussion/justification.

- Sections 2 and 3 are partial (2) or missing (3), I guess what I read is just a preliminary version;

- I was quite puzzled by reading Section 4. I understood that the knowledge of R is a pre-requisite for reading the book, but it seems strange to me to get an R output introduced by "The R code below provides more info about the contents of the dataset, values of the variables, etc." without any further explanation. The feeling is: "here the code, read it by yourself..."; similarly, the R output of the models (logistic regression, random forests and gradient boosting) is there without any explanation: what does it add? I guess a description of the results is necessary, otherwise it may be more reasonable to remove everything. The current solution is very puzzling;

- in the same section, since the authors, in my opinion correctly, give an insight on the data, I was expecting some information on the correlation structure, as its importance is pointed out by the authors when they describe the pro and cons of the various visualization tools. On the same line, it seems strange to only have univariate graphical analyses for data exploration. Again, I think the choice is between doing a proper analysis or not doing anything, this "partial solution" may lead to bad practice when the reader will analyse the data herself/himself. And I would like to stress again the feeling of being left "alone" with the outputs, without any comment (e.g., how should I interpret the violin plots? What do they say to me on this specific dataset?). I understand that this step of the analysis is not the focus of the book, but, if something is done, having these plots without explanation is quite strange...

- I was also wondering why gradient boosting has only been used on the first dataset and not in the other two. Maybe it will be clear in the following chapters, but at the moment one (at least I) does not understand why these specific methods (logistic/linear/multinomial regression and random forests) have been chosen, and why boosting only in the first dataset...

- As a final comment on section 4, I think more carefulness should be used in the presentation of the plots (considering that the book focuses on visualization): for example, the values on the y-axis of Figures 4.3, 4.4 and 4.5 are not really readable.

- I think that the discussion on what "instance-level" and "global level" are should be done in Section 1 when these terms are introduced. I do not think the reader should wait for section 5 to understand that;

- I suggest being consistent among the sections (but probably all sections from 5 to 9 can be collapsed in a single chapter) about the presentation of the different plots. The aspect I suspect the reader may be interested the most in, the interpretation of the plot, is in a case (section 6) contained in the subsection "method", in other cases (section 7 and 8) in "intuition";

- I also found strange the part where the intuition of the CP plot is described in comparison to a method, LIME, that at this point one is not supposed to know yet... If the reader does not know LIME in advance, (s)he gets easily lost;

- there are parts, like on page 72, in which the language is very technical, and seems to require a deep knowledge of the package ggplot2;

- I suggest being careful in the interpretation of the examples. E.g., the last sentence of Section 7.4 must be clearly linked to Figure 7.1. As it is now, it seems that one can get this information only by looking at Figure 7.2.

All in all, I think that the book addresses important issues and can be useful for practitioners. From what I read, nevertheless, I would ask the authors to be more methodical in their exposition, and carefully review the structure and the content of their chapters.

Finally, I did not look specifically for typos, but I, nonetheless, identified a few:

- page 12, line 1: forumlated -> formulated;

- page 14, line -5 (5 from the bottom): mmay <- may;

- page 14, line -4: "which features and how influence" <- "which and how features influence";

- page 16: line -10: wth -> with;

- page 17, line 9: one cannot define a function from a space to a point, it should be $f: \mathcal{X} \rightarrow \mathcal{R}$ (or whatever symbol to denote the space of real numbers);

- page 43, line 13: apatments -> apartments.

Review 2, chs 1, 3-9

This book touches on a very popular and important topic of predictive modeling. The book may serve both as a reference or textbook. It is well written with each section having unified structure. Subsections containing intuition concepts are very much appreciated.

Specific comments:

1. It appears that the title "Predictive Models: Visualizing, Exploring and Explaining: With Examples in R and Python" would be preferable over currently proposed "Predictive Models: Visual Exploration, Explanation and Debugging". In particular, debugging part of the title is difficult to accept.

2. In addition to already included three examples of data, it would be very helpful to have an example of data for predicting a 5 or 10-year survival after a specific event, for example after cancer diagnosis.

3, Consider moving sections 4.1.7, 4.2.5, 4.3.5 to appendix. In this reviewer's view these sections do not pertain to predictive modeling per se and to some extent they distract from the flow of the book,

Review 3, chs. 1,3-9

1. Who would find this type of book useful? Can you describe a kind of book that is needed in this area? Will the book serve as a reference, textbook, or both?

   The book is very interesting and will be used by data analysts (not just statisticians) who are interested in prediction. It can be served as a textbook in class for graduate students or as reference book for data analysts and researchers who are using predictive models for their analysis.

2. Would you recommend any changes in the contents that would make this book more useful?

Please see my general comment about R code in question 5. Below you can find few comments that I wrote while reading the text.

Page 13:

Use italic font for white and black box, i.e. \textit{white box} instead of the quates.

Page 13:

introduce the melanoma example before you discuss Figure 1.2

Chap 4 & Chap 9: The two chapters include an abstract (page 26 and page 95). These abstracts do not appears in other chapters.

Page 27:

Variables names in page 27 are printed in italic font, it is better to present them in an R font i.e., \texttt{gender} instead of \textit{gender etc. In that way, the variable name will be the same (i.e. the variable names in the output in page 29 will be the same as the variable names in page 27). The same comment for the variable names in page 43

Page 29:

try to avoid a long R output which is not explained and which is not needed. For example, the output in page 29 is explained in details in page 27 so it is not really add anything. Instead, you can provide an online program that produces the results presented in each chapter so readers can reproduce the same results in their computers.

Page 30:

Try to include the R code and output as a part of the text for example, in page 30 you can present the R code as a part of the text in the following way (your original text is in red) .

In particular, we replace the missing age
values by the mean of the observed ones, i.e., 30. Missing
country will be coded by “X”.

```
titanic$age[is.na(titanic$age)] = 30
titanic$country <- as.character(titanic$country)
titanic$country[is.na(titanic$country)] = "X"
titanic$country <- factor(titanic$country)
```

For sibsp and parch, we replace the missing values by the most frequently observed value, i.e., 0.

```
titanic$sibsp[is.na(titanic$sibsp)] = 0
titanic$parch[is.na(titanic$parch)] = 0
```

Finally, for fare , we use the mean fare for a given
Class , i.e., 0 pounds for crew, 89 pounds for
the 1st, 22 pounds for the 2nd, and 13 pounds for the 3rd class.

```
titanic$fare[is.na(titanic$fare) & titanic$class == "1st"] = 89
titanic$fare[is.na(titanic$fare) & titanic$class == "2nd"] = 22
titanic$fare[is.na(titanic$fare) & titanic$class == "3rd"] = 13
```

In this way, you can avoid a long parts of R code and make the R code an integral part f the text.

Page 34:

Variable names are in italic (survival) and in R font (fare) use the same font (preferably R font \texttt{fare) etc.)

The output of the logistic regression model is not discussed in the text, the same for the random forest models, for example what are the confusion matrices presented in the output mean etc?

Page 39:

The characteristics (age class etc) and the predicted probabilities for “Henry” are not discussed in the text, “Henry” should be presented in the text similar to johny_d in page 38.

Page 44-45:

The R code for the figures presented in Section 4.2.1 is not available. For completeness, it will be useful to include such a code. For example for Figure 4.16 (your original text is in red) you can use:

Figure 4.16 , produced using the following R code,

R CODE FOR THE FIGURE

indicates that the highest prices per meter-squared are observed in Srodmiescie (Downtown).

The same comment for Figure 4.17-4.19 presented in Section 4.3.1

Page 45:

Figure ??

Page 63:

The point in Figure 6.1 B is blue, not black as mentioned in the text.

Page 70:

It is not clear what is presented in the cp_titanic_rf object.

Page 81:

The code presented in page 81 (section 7.6) is presented before in page 69 and also I page 91 and the same for the code of the object cp_titanic_rf

Page 82:

The figure is presented before in page 80

Page 83:

It is not clear why you replace “1” (as in page 82) with Henry in the R code

oscillations_uniform$`_ids_` <- "Henry (this R code is also presented in page 82)

.

2. Please explain why you do or do not regard the manuscript as technically correct, clearly written, and at an appropriate level of difficulty. What are its strengths and weaknesses? You may comment on the manuscript.

Up to page page 65 (section 6.5) technical details are not an issue. The assumption is that predictive models based on either logistic regression or random forest are known for the reader (the same for the Multinomial logistic regression in Section 4.3.2). To make this book more accessible for readers, it will be useful if these topics are presented/discussed in an appendix so readers who are less familiar with these topics can read it there. This is especially important if you developed this book for graduate students (in stat or data science) who are less familiar with these topics. You can also include basic examples in this appendix so the student will be able to train themselves on these topics. It will also help non-statisticians to read the book.

4. What other books are available on this subject? Do they have any particularly strong or weak features? Does this book offer any significant advantages?

There are books that focus on prediction models, for example the element of statistical learning and an introduction to statistical learning but these are not focused on the evaluation of predictive models which is the main focus on the proposed book and its main advantage. As predictive models become very popular in the last years, such a book that focus on the evaluation of the models and model diagnostics can be very popular.

5. Please explain why you would or would not recommend publication. If you would,

that are the most important changes that should be made before publication?

The main change that can be done is to include the R code/output as a part of the text and not to present relatively long outputs. Also, if you provide an R program per chapter, the readers will be able to re produce the results presented in the book and this will help them to nderstand/implement the methods. This will be extremely useful if the book will be use as a text book in graduate master programs. It will also help readers who are less familiar with R to run the analysis.