The Final Report

Big Data Analytics: How to Write a Great Final Report

The final report of a project explains what you did, explains who will be interested in this analysis and how they can use it, presents your results, and explains the business (or other decision-maker) significance of your results. It should be thorough enough that a reader who as your initial data can reproduce your results, as well as understand why you did your analysis in a certain way. The detailed or routine parts of your report, such as your R code and full versions of graphs and tables, should go in appendices.

Each report serves at least three audiences:

1. A hypothetical decision maker who will use your analysis to decide on actions.
2. Academics, both your professor and future students who will learn by figuring out what you did.
3. Prospective employers, who you will try to impress with your machine learning abilities.
4. Sometimes, actual real-world decision makers who will alter their actions because of the insights you provide.

This document contains:

- An approximate table of contents for reports
- A checklist of items to remember
- Discussion of how to measure and explain your results
- The grading template I have used in the past to grade final reports.

Prof. Roger Bohn Rbohn@ucsd.edu

May 18, 2018 Version 1.2

Project table of contents

Term papers - approximate table of contents

Criteria for final papers in *Big Data Analytics*. Version 1.3 May 18, 2018

These are suggested sections for the reports. Not all reports will need all of these sections, because each report is different. You may also decide to sequence your report differently.

This is also the outline I will use for part of my report grading. Grades will include "degree of difficulty," how well that section of work was done, and how well the work was explained and presented in your report.

1.1.1 Executive summary

Remember that executive summaries are quite different than abstracts.

1.1.1.1 What will someone learn if they read this report? Motivate

1.1.1.2 Sketch the problem, and why it's relevant/interesting.

1.1.1.3 Summarize your key results. (Even if they are negative: "there are no useful patterns here.")

1.1.2 Introduction: Understanding the problem you are investigating

1.1.2.1 What is the domain (business/industry/set of issues)? *Example: Short reviews posted on e-commerce sites.*

1.1.2.2 What is the question/problem.

a. *Example: How many reviews for shoes are fake? Is there an easy way for consumers to guess at the fake reviews?*

b. Example 2: How do patterns of Medicare spending vary by region of the country. (Using the new Medicare data release.)

1.1.2.3 Why is it interesting? Who cares? Implications? There is an existing literature on the subject; how does the new data contribute?

1.1.2.5 What are underlying theories about relationships you expect

1.1.2.6 Literature review: who has looked at similar data and questions. What did they find?

Hint: Never claim "this is the first study like it." Maybe it's the first time an issue has been studied with your data set, but it is never the first time a related topic has been addressed somewhere.

1.1.3 Scraping or creating the data (if needed)

1.1.3.1 Most of this discussion goes into an Appendix. Explain how much work you

did here, show samples of the code, show raw source pages and what the data looked like before and after you worked on it.

1.1.3.2 Although this is not so important in most "Real" papers, it gives you grading credit for hard work!

1.1.4 Understanding the data, and making it useful.

1.1.4.1 Where is the data from? Who created it, what is the unit of analysis, etc. (Show useful samples or screen shots in an appendix)

1.1.4.2 What does the data set cover?

- How many observations? If you are not using the whole thing, what part are you using (and why)?
- What was the sample? What biases are likely because of who was and was not in the sample (external validity of your results)
- Be very clear about definitions of variables (e.g., are they per year? per customer? per month?)
- Any significant problems with missing data, erroneous data, multicollinearity, etc. How you dealt with them.
- Statistical descriptions - details can go in an appendix, with a summary table in the main text.

1.1.2.4 Transforming the raw data to make it more relevant (e.g. converting event times into durations).

1.1.4.3 For variables: Meanings, units, measurement methods, magnitudes. Measurement methods are important because they can be critical to what information you can, and cannot, extract, and how you interpret the results. Don't settle for superficial descriptions - be sure to get samples of the actual questionaires, actual texts, etc.

1.1.4.4 The role of *time*. Who are the decision makers in the world, and what data is available to them at the time they make decisions? Show the time sequence of information arrival and decisions by different parties, such as borrowers and lenders.

1.1.4.5 Exploratory analysis: basic box plots, cross-tabs/scatter plots, whatever is appropriate for understanding
***this* data set. Use some color where appropriate.**

a. Tutorial on box plots: http://www.r-bloggers.com/box-plot-with-r-tutorial/

1.1.4.6 Data cleaning. Much of this can go in an appendix.

a. *Examples: character versus numeric; categories;*

b. Discuss missing values.

c. Outliers and apparent errors in creating the data.

d. Translating into machine readable forms, e.g. separator values, expressing large and small numbers, negative numbers, etc.

1.1.4.7 Physical interpretations. For example is a measurement a flow (a rate over time), or a stock (a snapshot at one moment). Translate to dollars wherever relevant.

1.1.4.8 Transformations: Creating new variables out of the old, e.g. changing dimensions to area or volume;
changing levels of the same variable at 2 times to a rate of change, etc.

a. *Scaling the variables as required for your algorithm (eg. nearest neighbors kNN depends on scaling).*

b. *Raw data is timestamps; you convert that to the duration of events and time between events.*

c. *Example 3: Given variables for height and length of a piece of land, create a new variable for the area of that land.*

1.1.4.9 Summarize the data after transforming. Box plots, statistical summaries, crosstabs/scatters, etc.

▾1.1.5 Selecting and using data mining algorithms, and variables

1.1.5.1 Discuss choice of methods.
Do not present this information chronologically; instead present your ultimate solution. If you tried 4 very models and 40 different tunings of the models, empHasize the one or few that you ultimately decided were best. These may be the *last* ones you tried.

1.1.5.2 Discuss choice of error functions. Root mean squared error is only one of many error functions.

1.1.5.3 Trying different parameters for best results. Tuning the model.

1.1.5.4 Selecting the variables.

1.1.5.5 Measuring performance of model using multiple measurements

1.1.5.6 Using a validation data set to test model parameters, such as how many variables.

1.1.5.7 Using a separate test data set to estimate how it would perform in actual use.

a. You may have to think about the role of time in selecting the data subsets.

1.1.5.8 Compare to models others have used with similar data

▾1.1.6 Reporting quantitative conclusions

1.1.6.1 Robustness checks: what if you change the model slightly?

1.1.6.2 Other checks on data mining. Especially out-of-sample error measures of various kinds.

1.1.6.3 Magnitudes: *How big* are the effects? (This is *not* the same as statistical significance of a hypothesis test.)

1.1.6.4 What are the effects of these results on the original problem? For example, how much can you improve costs, or sales, or health?

1.1.5.9 Business implications of the model results should drive the way you report results. *Example: lift curve translated into $.*

1.1.6.5 Be thorough but emphasize what is important.

a. For example, if you have 20 variables in the model, report some results for all 20, but look carefully only at the ones that "matter," using appropriate measures of what is important in both business terms and BDA terms.

1.1.6.6 Proper reporting tables and diagrams. NOT a data dump; highlight the key points with arrows, circles,
bold, etc.

▾1.1.7 Reporting overall (managerial) conclusions. IMPORTANT

1.1.7.1 Translate from ANALYTIC results into business (or problem) relevant measurements. Use approximations and back-of-the-envelope calculations where necessary.

1.1.7.2 What have you learned?

1.1.7.3 Summarize the conclusions in words, graphs, and tables.

1.1.7.4 What are the *key* results?

1.1.7.5 "So what" - why should anyone care? If you cannot decide, you probably need to do more analysis.

1.1.7.6 What is the significance of these results, for different communities?

a. The original business or site that the data comes from

b. The people (such as consumers) being studied

c. Governments, regulators, etc.

d. How some group should change its *behavior* based on this insight.

1.1.7.7 How do your results fit with/extend/contradict the existing literature.

▾Limitations, unanswered questions, further research, etc.

- Anomalies - results that seem inconsistent with each other or with other studies. These can be a clue to important insights, but you may not have time to pursue them.
- You may have worked with a subset of the available data for some reasons
- and many more

1.1.7.8 Next steps if someone were to pursue this. What are interesting ideas that you did not have time for?

Appendices: detailed supporting data needed to fully understand your methods

- Some of these can be done either as an exhibit, or an appendix.
- Samples of the raw data (eg. an Amazon review page)
- Full description of the variables.
  - Name used in the report
  - Verbal description
  - Distribution of the variable in your data set, such as the mean and sigma. Boxplot diagrams.
- Exploratory data analysis, especially where it is relevant.
- R or other code used for the *final* model. Skip anything that you later threw away

Footnotes or endnotes, including references

- I don't care what precise reference format you use, but it should leave no doubt about the exact source, including the title and where it was published. A URL by itself is not enough.
- Remember that the more impressive the place of publication, the more credible the reference is, and by extension the more credible your paper. For example if something appeared as a working paper and in a good academic journal, cite the journal version.
- As a courtesy, include a URL or DOI number so the reader can find it. (If possible.)
- If your footnotes/endnotes are thorough, the references at the end can be limited to a few key ones.

Report Checklist

Double-check these issues

Write this paper as a Big Data Analytics paper, *not* a classical regression analysis.

a. Our goals are prediction and creating valuable insights, not testing a theoretical hypothesis.

b. Avoid the language of hypothesis testing, statistical significance, etc. Instead use the BDA equivalents.

c. Report standard data mining measures of model performance.

d. Re-read the Hal Varian paper I assigned if this is counter-intuitive. He discusses how to bridge the gap.

The paper must be readable by someone who has never talked to you in person, and does not know what you did.

a. Too often, papers leave out key facts, such as whether the data comes from humans or animals, or whether there is one observation per customer or one per item purchased.

Have you subtly called attention to how hard you worked?

a. Sometimes 70% of the effort in a BDa project can go into up-front work such as data cleaning.

b. A faculty reader may miss this, and wonder why you seemingly did little.

c. One solution is to build up appendices.

d. Another solution is to include a "work log" from the project as you proceed. Then use bar charts to show graphically where your time went.

What are weak links in your analysis? If someone wanted to dismiss your results, what would they point to?

a. Then improve those areas.

Look for good graphics which illuminate your analysis.

Suggestion: have an intriguing figure on your cover page.

All tables and figures should be fully explained within themselves (without having to read the body of the text)

Captions, definitions, which direction is "good," whether a "1" means male or female, etc.

Appendices

1.1.9.1 Samples of the underlying data. Very important - it will answer many

questions if you have left something out of your descriptions.

1.1.9.2 Discussion of data scraping and conversions

1.1.9.3 Details of various printouts

1.1.9.4 R code. REQUIRED

1.1.9.5 Bibliography. In proper format with name and source of each paper. URLs where possible.

Data Analytics FAILs to avoid

Messing these up won't literally lead to a failing grade by themselves, but they are all required unless you have a *really* good and explicit justification.

1. Use hold-out data to evaluate the accuracy of your model.
    - Best if you use both a validation set (to calibrate and tune your models) and a final testing set.
2. At least one of your analyses must include an explicit prediction or classification.
    - Your core analysis must be for making decisions, rather than testing a hypothesis.
3. Do not use R^2, t statistics, etc.
4. Classification results must be reported with one or more confusion matrices.

Measuring results in projects

What is your outcome measure? Justify it in business/ physical terms.

See the forum discussion "Including cost of errors in your analysis."

> No predictions are perfect, obviously. One objective in data analytics is to minimize the cost of errors. Inclassification problems, this cost is the sum of costs of False Positive and False Negatives. In continuous predictionproblems, the cost is some function of the size of the errors. The purpose of this discussion is to help each team figure out how to approximate the cost of errors in its project. Then use this cost to partially guide your analysis, and report your results in terms of costs incurred.
> **Classification problems** (binary) are easier to talk about. Typically you are deciding on an action for certain people, and the cost is the sum of: [cost of actions for predicted positives] + [cost of errors for false negatives]. An example was the mail order catalog classification example that we did: cost of mailing catalogs + cost of lost sales to people we did not mail catalogs to.

Example: One team is looking at the hospital readmission rate. If a patient gets readmitted within 30 days, treatment is not reimbursed. That cost varies, of course, but suppose the average is $10K. Patients who are judged to be high risk of readmission can receive special counseling and follow-up from the hospital. Suppose that costs $1000, for 4 home visits by a nurse at $200 each, plus an expected value of $200 for patients who need to be brought back for special treatment. So the **critical ratio** is 1:10, the cost of a predicted positive (whether the prediction was true or false does not matter) versus a false negative. (This is similar to a **critical fractile** calculations in TOM.)

Include discussion of the cost of errors in your final report. Use those costs to determine where on the ROC curve you should position your solution. In other words, where to set the cutoff on your classification algorithm. I can provide advice to each team on how to estimate these costs, but first give your preliminary ideas about what costs should be considered, even if you are not sure how to put numbers on them. That way, the entire class can learn from the discussions. Of course, estimates should ideally be based on reading articles/academic papers about similar situations.

Prediction (continuous variable) problems: The costs usually depend on the size of the error. Although classical statistics assumes the costs are quadratic in either direction, in real life they are usually determined by the economics of the business. For example, suppose you are predicting demand for products at a retailer. Then if your forecast is too low, the relevant cost is (cost per stockout) x (forecast error). If your forecast is too high, cost is

(amount of excess inventory) x (some kind of carrying cost per unit time).

Grading Template

Grading Template for Big Data Analytics, Final Report

May 2018 Roger Bohn version 2.5

Paper Title: **Authors**

1 = Topic not mentioned 2 = Mentioned, but little useful work done.
3 = Decent discussion 4 = Good discussion

Introduction	Score 1-4	Comments
Motivation for the problem		
Summarize the key results		
** Understanding the substance of the problem and its domain (business issue etc.)		
Use of outside literature on related problems		
Problem solving insight		
Is the problem/domain understood well		
Good exploratory analysis - what is in the original data, what patterns?		
Interpreting the results - how big are they in managerial terms, what do they say about what to do, how conclusive?		
Scraping creating +cleaning data		
Acquiring data, scraping etc. Beyond Kaggle data?		
Data cleaning, data filtering		
Merging data sets from diff. sources.		
Describing & Exploring the data		
Basics: sources, #observations, level of each observation,		
Describing variables: text, distribution info.		
Role played by time in the data		
Scatter plots, cross-tabs, etc.		
****Model formulation/feature engineering**		
Dependent variable: definition, formula, distribution. Log or other transforms. Several versions		
Feature construction, going beyond original variables		
Conceptual cleverness, validity		
Multiple models/studies from same data		

Solving the model		
Using mining methods appropriately (eg. interactions, normalizing variables)		
Using multiple solution algorithms (linear + nonlin)		
Interpreting the results eg financial impacts.		
Going beyond the initial results.		
**Reworking and tuning model to improve results		
Clarity of analysis		
How clear is it what you did?		
Is the problem clear?		
Are the results clear? Appropriate ties between text and numbers/appendices/tables?		
** Clear implications. Interpretation; Conceptual insights		
Overall writing style.Clear sentence structure etc.		
Presentation of work, including technical work		
Charts and figures - readable, well chosen types, thoroughly cleaned up (e.g. axis labels, colors, legend).		
Important information shown - e.g.s distributions of key variables, meaningful values of key coefficients?		
Is this a Big Data Analytics paper, or are the methods and write-up simply like a standard regression paper? (QM3)		
Word equations, figures, tables, versus raw output dumps		
Mechanics of writing		
Checklist for final reports: file name, date, references, page numbers, etc. See separate checklist.		
Excellent Spelling and grammar, clear sentences.		
Adequate footnotes. Fully documented so a reader can find them.		
Avoiding bugs (e.g readable numbers, units))		

** = Special importance

Degree of difficulty

Projects have many stages, and doing difficult work at any stage receives credit. Scoring is somewhat like Olympic diving, where dives are assigned a "degree of difficulty." Interpret the following issues as:

High credit —— Medium credit — Minimal credit

- Text mining of long documents —— Short documents —— No text
    - Extra difficulty for bigrams, Latent Semantic Analysis or similar, mood analysis or other work beyond bag-of-words etc.

- Scraping from a web site over several weeks — Scraping an existing web site — Using data that someone else cleaned, simplified, and documented e.g. from Kaggle
- Transforming data heavily, e.g.:
    - using time stamps to infer what happened between direct observations
    - Logs, ratios, financial formulas
    - See section "model formulation/feature engineering"
- Business or technical modeling of a process. — Using numeric variables without attaching any meanings to them.

Discussion: Things done well

Things needing more work

Modeling and technical issues

What's required to get a working paper from this report?