The Do's and Don'ts of Data Mining [slideshare]

By: Heather Hinman, Salford Systems
February 27, 2014

Since the dawn of data mining (beginning with Bayes Theorem in the 1700s), there have been many successes and failures, even by the top experts in the field. No matter what job function or industry you work in, it is generally agreed on that on-the-job training is a far better learning tool than any classroom lecture. Learning from our mistakes is one of the ways we move forward and accomplish our goals. The same goes for data mining practitioners and data scientists; hands-on experience (or lack thereof) results in victories and blunders that set the foundation for advancements in the field.

I asked a few data scientists I know to offer examples of their own do's and don'ts from their real-world experience working with data, consulting, product development, and professionalism in the field. I had an OVERWHELMING response from these kind contributors and have included some of them in the SlideShare below. Don't forget to share them with your colleagues who are committing the DON'TS of data mining!

View these on SlideShare

The Do's of Data Mining

1. **Do plan for data to be messy.** While data is available for mining projects in ever-increasing amounts, it is the rare occasion when it will arrive in a tidy, mining-ready format. More typically, it will show up in multiple spreadsheets that vary in format and granularity. These varied formats frequently require hours (and hours) of ETL (Extract, Transform, Load) time.
2. **Do create a clearly-defined, measurable objective for every project.** While most objectives are good at stating what you want to achieve, it's not the "what" that matters. How will you determine project success? And by when? These three elements...the what, the how, and the when... are not options, they are requirements for a clearly-defined project objective.
3. **Do ask questions.** Understanding the problem and asking the right question is more important than using an advanced algorithm.
4. **Do simplify the solution to increase your chances of success**. Most Data Scientists are aware that simpler solutions are generally better solutions. Why? Because they have fewer moving parts that can break and there's less likelihood of model overfitting. When simplifying, consider the predictors, but also the target variable... ask, "Can it be simplified as well?"
5. **Do cross-check data coming out of the ETL process with the original values, and with project stakeholders.** This is the time to uncover errors in the ETL process or in the raw data, itself. Stakeholder data reviews are critical for making sure everyone agrees that the proper data is being used. Waiting to uncover anomalies and errors during mining/modeling is too late and wastes everyone's time. Plots and descriptive statistics can be helpful in spotting issues.

6. **Do use more than one technique/algorithm.** Given the availability of tools like SPM and others, "it takes too much time to try multiple techniques" is no longer a valid response. For example, if the problem is one of classification, don't just use CART and say "that's the answer." Random Forests may deliver better results.
7. **Do be informed.** Stay fluent on the latest data mining concepts and approaches, as well as data mining history.

The Don'ts of Data Mining

1. **Do Not Ever … I mean <u>EVER</u> underestimate the power of good data preparation.** THE number one mistake that Modelers make is related to lousy or totally absent data preparation prior to model development. Good data prep includes cleaning, transforming, and aggregating model input data as well as the identification and treatment of outliers.
2. **Don't use the default model accuracy metric.** The default metric for continuous valued prediction is R-squared or the average squared error or mean squared error (MSE). The default metric for classification problems is percent correct classification (PCC). These are both batch metrics that summarize the accuracy of your models in a single number. While some business problems should assess model accuracy using MSE or PCC, these are rare in my experience. Most often, some errors are worse than others and we select models that do more than have good accuracy *on average*. They are the best at selecting subsets of the population for treatment.
3. **Don't forget to document all modeling steps and underlying data!**
4. **Don't overfit**...with Big Data it is easy to find patterns even in random data. Use appropriate tests such as randomization tests to avoid finding false patterns in test data, which will not hold later on.
5. **Do not just collect a pile of data and "toss it into the big data mining engine" to see what comes out.** Domain knowledge is an important cross-check on the variables being used. Extraneous data can reduce model accuracy.
6. **Do not ascribe them mystical powers and wrongly think "it's all about the algorithms".** Don't overly-focus on software, make unfair assumptions about innate algorithmic capabilities, or even the cool software buttons that can be pushed. All are ill-equipped and are poor substitutes for your intelligence, insight, skills, and creative abilities.
7. **Do not underestimate the power of a simpler-to-understand solution that is slightly less accurate.** A model a client cannot grasp is one that will not be trusted as much as one that "makes sense."
8. **Do not blindly trust assumptions made to satisfy frequency statistics, as well as p-values and AIC.**

Thank you to my expert contributors!

Scott Terry, President of Rapid Progress Marketing and Modeling, LLC

Scott is a multi-talented industry veteran with 30 years of experience on both the client and services side of data mining, direct and database marketing. He's lived in your shoes and knows how to turn issues into opportunities. *(Do #2,#4; Don't #1,#6)* Website: www.RPMSquared.com

Dean Abbott, President of Abbott Analytics

Dean has over 21 years of experience applying advanced data mining, data preparation, and data visualization methods in real-world data intensive problems, including fraud detection, response modeling, survey analysis, planned giving, predictive toxicology, signal process, and missile guidance. *(Don't #2)* Website: http://www.abbottanalytics.com/

Gregory Piatetsky-Shapiro, Editor for www.KDnuggets.com

Gregory is an analytics, Big Data, data mining, and data science expert. He is a KDD & SIGKDD co-founder, part-time philosopher, and dad. *(Do #3; Don't #4)* Website: http://www.KDnuggets.com

Jim Kenyon, Director of IT Services for Optimization Group

Jim has held management roles with a number of tech firms including ADP, Just Talk and Gale Research. A published author, Jim is fascinated with the methodological application of technology to solve business and scientific problems. *(Do #1,#5,#6; Don't #5,#7)*
Website: http://www.optimizationgroup.com/

Falk Huettmann, Wildlife Ecologist for the Institute of Arctic Biology

Falk's work is explicit in space and time, and looks closely at the global effects of the economy. His research interests include: wildlife ecology, seabirds, predictive GIS modeling, web-based wildlife databases and metadata, spatial aspects of Population Viability Analysis (PVA), landscape ecology, Russian Far East, tropical ecology, and conservation steady state economy. *(Do #7; Don't #3,#8)*
Website: http://faculty.iab.uaf.edu/