Article from

Predictive Analytics and Futurism

December 2015
Issue 12

Getting Started in Predictive Analytics: Books and Courses

By Mary Pat Campbell

Back in September 2009, this section sported a brand new name: the Forecasting & Futurism Section (before it had been the Futurism Section). In the inaugural newsletter that month, introducing the new name, there was also an article introducing Forecasting concepts: "Introduction to Forecasting Methods for Actuaries" by Alan Mills. Alan put together a handy table listing common forecasting approaches in actuarial work, as well as references for those methods.

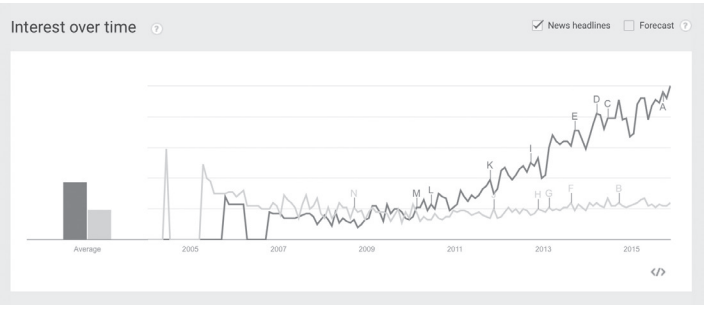
At the time, "Predictive modeling" was relatively new, and he noted it was gaining in popularity.

Here is how Alan described the method:

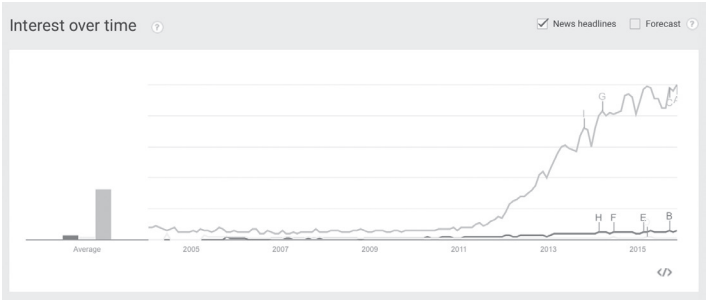
> "An area of statistical analysis and data mining, that deals with extracting information from data and using it to predict future behavior patterns or other results. A predictive model is made up of a number of predictors, variables that are likely to influence future behavior."[1]

Since that overview article from six years ago, predictive modeling and analytics have taken off—so much so, it's now part of the name for the section!

"Predictive analytics" and "Predictive modeling," have caught on broadly, and in insurance, first being particularly used in property & casualty pricing applications. "Big data" has really risen in popularity as a search term since 2012 … perhaps partly due to the prominence of people like Nate Silver of 538 fame.



"Data Source: Google Trends (*www.google.com/trends*)." See *https://support.google.com/trends/answer/4365538?hl=en.*



"Data Source: Google Trends (www.google.com/trends)." See https://support.google.com/trends/answer/4365538?hl=en.

Actuaries have the ability to pick up predictive analytics concepts—some of which are not very complicated at all, just being linear regression models from large data sets. But predictive analytics goes beyond Generalized Linear Models, and even with GLMs there are niceties that actuaries should know about.

BUT WHERE TO BEGIN?

Below are some resources for the beginner in predictive analytics … and sometimes a nice way for those already well-versed in many of the techniques to expand to a few more they had not considered.

There are two main threads involved in getting started with predictive analytics:

1. Statistical theory and modeling—understanding the approaches, what each does, and what the strengths and weaknesses are for these; and

2. Computing—specialized software and languages intended for crunching Big Data and performing analytics.

I am going to try to pick resources that combine the two, but sometimes that is not possible. For the most part, I will be highlighting free or inexpensive resources.

BOOKS

STATISTICS (THE EASIER WAY) WITH R BY NICOLE RADZIWILL

Weblinks	Free preview: https://qualityandinnovation.files.wordpress.com/2015/04/radziwill_statisticseasierwithr_preview.pdf
	Amazon link for book: http://amzn.to/1URjyQD
Languages/Topics	R and Introductory statistics—confidence intervals, regression, and statistical inference
Level	Absolute beginner

I partly picked this book because the author is a long-time friend, but also because this is a very easy entry into using R as well as thinking about statistical models. The statistics material in the text

is similar to the syllabus of the Statistics VEE, so the topics should be familiar to actuaries.

R is a free statistical software package, and thus is used in many of the predictive modeling texts one finds. However, most statistics texts using R have a large gap in explaining how one uses R … and most R texts have a large gap in explaining the statistics while walking you through how to use R.

Nicole developed this text through her own classes at James Madison University in Virginia (Dr. Radziwill is an assistant professor in Integrated Science and Technology at JMU) geared at undergraduate science majors. As Nicole writes, one of her target audiences was:

> "Smart, business-savvy people who want to do more data analysis and business analytics, but don't know where to start and don't want to invest hundreds or thousands of dollars on statistical software!"

I have gone back to Nicole's text as a reference for doing certain things in R, because she walks through every step. This book is long as a result of the step-by-step R code, but I have found this more helpful than trying to Google "how to do X in R."

DATA SCIENCE FROM SCRATCH: FIRST PRINCIPLES WITH PYTHON BY JOEL GRUS

Weblinks	Joel's site: *http://joelgrus.com/*
	Amazon link for book: *http://amzn.to/1URkqoA*
Languages/Topics	Overview of multiple data analysis techniques, Python, SQL
Level	Beginner

Python is another widely-used language in data analytics. While R was developed originally for statisticians, Python is a more general use programming language. That has led to differing groups of people developing already-created/written code for Python and R. Python is an extremely popular language due to its relative ease in use compared with other languages, and there have been several numerical computing packages developed for Python, such as numpy.[2]

Another disclosure: I am also friends with Joel Grus and previewed this text … I have a lot of friends. Joel is currently a software engineer at Google.

In this text, there is a quick introduction to Python—enough to run and adjust the code in the text. In addition to the linear regression and inference concepts that are also in the Statistics with R text previously, this text covers: clustering algorithms, Bayesian approaches, logistic regression, neural networks, and network analysis. He also covers SQL, because much of the data being used in the data-crunching first originated from SQL databases.

This text just gets you started in these techniques … in some cases, just enough to make you dangerous. While Joel does sometimes cover the pitfalls of certain techniques, his focus is primarily on how one executes certain types of analyses and not how they may go extremely wrong.

AN INTRODUCTION TO STATISTICAL LEARNING WITH APPLICATIONS IN R, BY GARETH JAMES, DANIELA WITTEN, TREVOR HASTIE, AND ROBERT TIBSHIRANI

Weblinks	Book's website: http://www-bcf.usc.edu/~gareth/ISL/book.html
	Amazon link: http://amzn.to/1URmvAL
	Online videos (free!): http://www.dataschool.io/15-hours-of-expert-machine-learning-videos/
Languages/Topics	More rigorous approach to statistical inference/modeling techniques, R
Level	Intermediate

For my last book recommendation, here is a more formal text (though "squishier" than the more advanced The Elements of Statistical Learning by a non-empty intersecting set of authors). It is more expensive than the two prior books, as this is a regular college text, and has the accompanying pricing.

That said: there is a complete set of online videos from a class based on this text. This will provide a link to the online courses I promote below.

I have been very slowly going through this text … the slowness due to me jumping back to other resources on R, so I make sure I understand what I'm doing. That's the weakness with this text—the R is not well-explained for the newbie. I would not start with this text for learning R, but once you've got a founding in R, the exercises in R are not so bad.

What's really nice is that you don't actually have to do any of the sections with R—if all you want are the concepts, you can skip the parts in R and pay attention to their worked-out examples.

Still, I think that doing the hands-on applied exercises in R is important in putting the pieces together.

As this is a "real" college textbook, it has end-of-chapter exercises, divided into "Concept exercises" and "Applied exercises." I really liked the "concept exercises" as they were geared to having the student probe that they really understand what is going on, and these exercises are very much geared towards thinking about which techniques are appropriate for which modeling tasks.

As an example, here is the question and my proposed answer for one of the conceptual exercises:

"4. You will now think of some real-life applications for statistical learning.

(a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

Classification may be useful if you're putting together policyholder data/response:

- underwriting in life insurance—have discrete u/w classes as opposed to more continuous u/w;
- might want to classify policyholders as being reactive/hot money vs. passive—very important in variable annuities; and
- might want to flag claims for possible fraud, but don't want to spend too much resources investigating every claim.

(b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

Regression useful in insurance:

- more continuous u/w as seen in auto coverages;
- if want to consider more continuous life u/w as with John Hancock's Fitbit program; and
- used in putting together projections of exposure in various p/c coverages. Can't observe everything while u/w, but may be able to find key variables.

(c) Describe three real-life applications in which cluster analysis might be useful.

- might be wanting to see if one can come up with new u/w buckets—cluster analysis may help;
- I used cluster analysis to see if there's common asset allocation strategies among life insurers—help tease apart influences; and
- could be used by exam committees to compare current exams against historical, check out various metrics (other than Euclidean) to see if there are clear outliers in exam performance" (COUGH COUGH CAS).

As I said, I've been working through this text, and my notes can be found at my dropbox: *https://www.dropbox.com/s/bf6dxtmtnzat4ny/Exercises%20from%20An%20Introduction%20to%20Statistical%20Learning.docx?dl=0* I have been trying to put in insurance/pension-related applications in answers to conceptual questions, but for some of the topics, it gets to be a bit challenging to think of actuarial applications … but give me time.

My R code for the book's applied exercises can be found here: *https://github.com/meepbobeep/ISLR*

Topics covered in this text: linear regression, classification, resampling/bootstrapping, model selection, dimension reduction in models, nonlinear models, tree-based methods (such as decision trees), support vector machines, unsupervised learning.

Disclosure: I am not friends with any of the authors of the following texts. Yet.

ONLINE COURSES

DATACAMP

Website	*https://www.datacamp.com/*
Languages/Topics	R and data science in general
Level	Absolute beginner to intermediate
Timing	On-demand, very short lessons
Paid features	Access to all courses, statement of completion
Credentials	Statement of completion

Datacamp has online lessons in R, which I originally found out about via a class on edX. Like many of the online courses below, they keep trying to upsell you. The pricing is by time period—either by month or by year (cancel any time!) I have tried only their free content, which tends to be the introductory classes. I suppose they figure if you get a taste, you'll want more.

I thought these lessons were very well-done, taking you step-by-step through R and some of the major tasks one would want to do in R when doing predicative analytics. However, the material I see on the site, even the paid courses, don't get to a very advanced level. However, it does touch on using R in ways the prior texts don't: for prettier graphs and dynamic reporting.

I found these lessons were very rapid to go through, and I'm thinking of paying for the one month of access ... should be no problem to go through 18 available courses over the Christmas break, right?

One of the nice features of the introductory courses is that you do not need to install R yourself—you will be able to run R code in the browser itself.

UDACITY

Website	*https://www.udacity.com/*
Languages/Topics	Data analysis, R, Python, SQL, Hadoop, (and much, much more)
Level	Beginner to advanced
Timing	On-demand, usually takes a few months for a full course (some mini-classes are shorter). Nanodegree and regular degree programs are on a schedule
Paid features	Monthly charge for access to coaches, projects with ongoing feedback, verified certificates and degrees (normal and nano-)
Credentials	Verified certificates, Georgia Tech MS in Computer Science, coming soon: nanodegrees

Udacity has a coding focus, along with applications such as with Front-end development and Data Analysis. For this review, I'm only looking at the courses in the Data Analysis nanodegree.

The classes on Udacity are more like regular classes, with quizzes and assignments. Udacity also has video lectures. Classes are rated for level, the advanced classes tend to have programming experience prerequisites. They have classes with serious Computer Science content, not only about how to program. They have classes built by various well-known tech companies, such as Facebook, Google, Amazon Web Services, Salesforce, and Twitter.

In addition to verified certificates for specific classes, and their partnership with Georgia Tech to provide an online-only M.S. in Computer Science, Udacity has recently created "nanodegrees" in specific areas, one of which is for data analysis. These nanodegrees are intended to be completed in less than a year. It looks like there was great demand, because they increased the fee for the nanodegrees from $150/month to $200/month in the past year, and have restricted enrollment in the nanodegrees to certain times of the year.

To access the classes for free, just click on "Start free course" on the specific class page. You can get to all the material: videos, text files, and even assignments. Within the videos themselves, they often stop for quizzes for immediate checking of understanding. Obviously, there are features you can't access if you aren't paying. The courses that are free are generally available on-demand.

The main place to start for their data analysis courses is Intro to Computer Science, which is mainly about learning to code

in Python. It seems most of their data analysis classes depend on Python.

COURSERA

Website	*https://www.coursera.org/*
Languages/Topics	So very much!
Level	Beginner to advanced
Timing	Most on specific schedules, 4-week to semester-long courses; a very few are on-demand
Paid features	Certifications (see below)
Credentials	Signature Track credential, Specialization certificates from sponsoring universities

I find Coursera the most dangerous of all the websites to go to … because there's so much there, and not all of it is programming. Looking at the list of stuff I've signed up for on this site: The Data Scientist's Toolbox, R Programming, Exploratory Data Analysis, Fundamentals of Music Theory, A Beginner's Guide to Irrational Behavior, Machine Learning, Introduction to Mathematical Thinking, Data Analysis, Comic Books and Graphic Novels, Computing for Data Analysis, An Introduction to Financial Accounting, Exploring Beethoven's Piano Sonatas, The Science of Gastronomy, Coding the Matrix: Linear Algebra through Computer Science Applications, Introduction to Data Science, and Gamification.

That's not necessarily exhaustive.

I obviously don't have enough time to seriously pursue all these courses, especially since, unlike the other sites listed above, most of these classes are built to specific time schedules, with classes starting and ending on particular dates. Usually, I'm only seriously following one class at any given time and downloading all the PDFs, videos, and other supporting documents … completely free. I have used some of the items I've come across to teach my own courses on other topics.

All of the courses on Coursera are backed by accredited institutions, and thus Coursera has a more academic feel than Udacity. Some of the classes come with paid certifications, and some courses have no free version at all. Many of the business-related data analytics courses are like that, I find.

Like Udacity, Coursera has developed something akin to "nanodegrees" called Specializations, which are short tracks of verified courses that take about a year to complete. A few of the Specialization tracks available as I write this article are Machine Learning (University of Washington, six courses), Big Data (UC-San Diego, six courses), Business Analytics (University of Pennsylvania, five courses).

Lots of courses to choose from at Coursera, and my main warning is to check prerequisites. Some of the numerical computing courses assume you know specific languages at particular levels. Some are truly introductory, and will walk you through how to get started in various languages, but many are at intermediate levels or higher for the coding, especially in the data analysis courses, so you want to be careful.

Got any favorite resources for the beginner in predictive modeling and data analytics? Let me know about them—*marypat.campbell@gmail.com.* ■

ENDNOTES

1 Alan Mills, "Introduction to Forecasting Methods for Actuaries." Forecasting & Futurism Newsletter, September 2009. pp 6-9. *https://soa.org/Library/Newsletters/Forecasting-Futurism/2009/September/ffn-2009-iss1.pdfn*

2 *http://www.numpy.org/*

Mary Pat Campbell, FSA, MAAA, PRM, is VP, insurance research at Conning in Hartford, Conn. She can be reached at *marypat.campbell@gmail.com.*