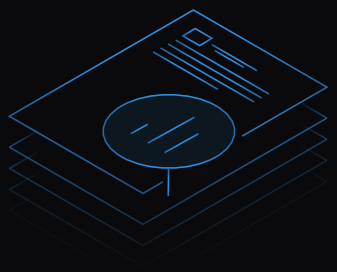
Teaching Predictive Analytics at the University of Colorado



Dr. Kai Larsen, Associate Professor of Information Systems
Leeds School of Business, University of Colorado

Predictive analytics is reshaping business and society, raising serious questions about how colleges anduniversities should prepare graduates. One answer may be to teach predictive analytics to all business school students. What would it take to implement this important vision and why is it not currently being done?

As a business analytics professor, Kai Larsen's goal is to teach a mixed range of students: those who immediately understand how predictive analytics has reshaped their future jobs (Information Management and Marketing), those for whom different flavors of business analytics have long since infused into the core of their fields (Operations Management and Finance), and those for whom predictive analytics currently is reshaping "only" a small part of their discipline (Accounting). It is becoming clear that all of these students must, at a minimum, understand predictive analytics conceptually to make decisions that will affect the future of their companies as machine learning tools continue to provide business insights and drive change within and outside the enterprise.

> This year marks a decade since I first started bringing predictive analytics into the business school classroom, sometimes with great success and other times markedly less so. Most of my failures were caused by abysmal tools and the massive number of manual steps needed to make up for the tool incapabilities, always pushing more content into an already bloated class.
>
> — Kai Larsen

"Only in the last year have I seen technological trends reversing. I now believe that with the right tools, for the first time, predictive analytics can reasonably be brought into the core of business school education. This is quite fortuitous, as predictive analytics is changing society at a dizzying speed. To send students into industry without at least conceptual knowledge of predictive analytics may be akin to sending students out without knowledge of accounting or marketing. In all these cases, students could probably survive for a while if they specialized properly outside of those areas, but they will clearly be unprepared for collaborative work and most leadership positions."

Kai Larsen

To be successful in teaching predictive analytics, professor Larsen sees two complexities that have to be addressed:

ALGORITHM-SPECIFIC PRE-PROCESSING OF DATA

For example, students must know that some algorithms such as regression will throw out a whole row of data if any predictor is missing a value. This brings with it a whole set of knowledge requirements for how to best impute those missing values.

ALGORITHM EVALUATION AND SELECTION

There are hundreds, if not thousands, of machine learning algorithms to select from. Some are commercially restricted but most are open source and available in specific packages in R and Python, or in special-purpose, cutting-edge packages like Tensorflow from Google. Predictive analytics used to require both an understanding of all of these algorithms and knowledge of how to select between them; the same algorithm will seldom be best for two different problems. As data sizes grow, evaluating all these algorithms brings outsized challenges for the infrastructure required to teach predictive analytics. These two complexities together explain why predictive analytics has remained in the purview of year-long MS programs in analytics and, so far, outside the core business curriculum.

But Before Modeling Can Begin, Certain Minimum Data Blending Skills are Required. This Presents Three Additional Challenges:

ACCESSING DATA FILES AND UNDERSTANDING THEIR FUNCTIONS

As anyone who has tried to teach R or Python in a core, required class can testify, teaching students how to consistently remember how to access comma-separated files — not to mention the sheer multitude of different functions required for Excel files, database tables, Twitter and social media feeds — is a major challenge.

JOINING DIFFERENT DATA TOGETHER

Generally speaking, this requires logic that derives from relational algebra, and presents a major challenge for most students the first time it is encountered.

AGGREGATING AND SUMMARIZING DATA

For example, if we want to analyze the likelihood of a customer switching cell phone provider the day their contract expires (churn), we may join their customer record with the table containing information on the five times they called customer service (data may exist on their level of satisfaction on a scale from 1-5). We would need to join the two tables and then aggregate the resulting table back down to the customer level by adding features related to their average, minimum, maximum, and final level of satisfaction.

Professor Larsen finds that attempting to equip students across all disciplines with predictive analytics skills has its challenges. Operationalizing predictive analytics is fraught with complexities, and getting data ready for predictive analytics in the first place comes with its own set of challenges.

Here's His Solution:

I've found success overcoming these challenges by bringing two tools into the classroom: DataRobot and Alteryx.

The conceptual goals of predictive analytics and its complexities can be addressed through the use of DataRobot. I believe that at this point, it is the only platform in the industry delivering automatic algorithm-specific preprocessing, understanding of algorithms, and automatic evaluation and selection of algorithms. While DataRobot provides advanced functionality that would likely stump and outperform most recent graduates of MS in Analytics programs, its surface characteristics are such that it could easily be used to teach predictive analytics to undergraduate students.

The DataRobot platform takes a rectangular matrix of data, and assumes that data blending has already taken place. The data is uploaded into the cloud environment, and DataRobot immediately goes to work on calculating basic statistics. The user is then asked to select the target variable and to click a big round "easy-button".

The system goes to work setting aside a random holdback sample of 20% of the data for eventual evaluation of the solution accuracy, splitting the data into folds for fivefold cross validation, characterizing the target variable, pre-processing the data in a number of different ways depending on class of algorithm, and running through a large sample of top-performing algorithms with 16% of the available data. The algorithms with their most common pre-processing step are then ranked in a leaderboard based on their performance on the fifth fold of the data (the part not used to build the model in one round of fivefold cross-validation). Once this round is over, the tool automatically selects algorithms to continue on with 32% of the data, and if, for example, one of the random forest algorithms outperformed other algorithms, then multiple versions of that algorithm are instantiated with different pre-processing steps. This is repeated before moving on to 64% of the data.

Throughout this process, the user gets to observe the gladiatorial sport of the different algorithms jockeying for supremacy. Finally, with 64% of the data used for training (16% for the final fold and 20% for the holdback sample), the best algorithms are "blended" together using four different auto-blenders which are themselves ranked in theleaderboard. The tool now allows use of the best algorithms to predict new cases, e.g., ("What is the probability that John, whose cell phone contract expires tomorrow, will actually take his business somewhere else?").

The models may be explored in terms of their key attributes with opportunities to discuss confusion matrices, precision and recall and area under the curve of ROCs, textual features driving the target variable as well as the most predictive features and their curves and patterns.

Simply put, my goal with DataRobot is to create two-hour analysts: workers who can take a rectangular matrix, conduct predictive analytics, and create a presentation for management, all in the space of two hours. Potentially, three such projects and their presentations can, under the best of circumstances, be performed in a day, each analysis comparable with what used to take months of effort by the most expensive business analytics experts. This is the conceptual part of the puzzle.

For data preparation, my favorite platform is Alteryx. It provides a workflow-based process with every tool you could ever wish for, ranging from data access and transformation to predictive analytics, geographic evaluations, and customer data access. I like it for its easygoing, quiet competence and ability to handle data ranging from small to truly large. The fact that Alteryx now comes with two DataRobot tools (the Alteryx Connector for DataRobot) – one for training a model and one for scoring new instances – should make enterprises take note. For the first time, there exists a set of data science platforms that makes A-Z analytics easy enough for any undergraduate to be ready to perform in line with all but the best data science competitors. Other data preparation environments that would complement DataRobot include Pandas, R, SQL, or Excel.

The value of DataRobot became very clear to me when I used it in my own research work. I had a very strongly technical PhD student work with me on one predictive analytics task for a project that took two to three months. Then we pulled the same data into DataRobot to compare results, and in one hour, DataRobot had outperformed the PhD student by a factor of two, simply because he had missed a class of algorithms that worked really well for thedata in question and had not thought to balance the training data. There is little doubt that predictive analytics will find a way into the core of most business schools. What remains is simply a question of which colleges will be first to empower their students with outsized advantages as they move into enterprises with many more targets of opportunity, as analytics-trained employees.

Dr. Kai Larsen is an associate professor of information systems within the Division of Management and Entrepreneurship at the Leeds School of Business, University of Colorado, with a courtesy appointment as an Associate Professor in Information Science in the College of Media, Communication and Information. He is a Faculty Fellow at the CU ATLAS Institute, and an affiliated faculty at the Silicon Flatirons Center in the CU School of Law. As Director of the federally supported Human Behavior Project, he is conducting research to create a transdisciplinary "backbone" for theoretical research. Dr. Larsen's research uses automatic text mining technologies to create an integrating framework forpredictors of human behavior. This research has implications for our understanding of all human behaviors, including technology utilization, investor decisions, and cancer prevention behaviors.

Contact Us

DataRobot
225 Franklin Street, 13th Floor
Boston, MA 02110, USA

www.datarobot.com
info@datarobot.com

powered by aws