Paper 1794-2018

Creating a Successful Data Science Program – A Joint Academic and Industry Perspective

Krzysztof Dzieciolowski
John Molson School of Business, Concordia University

ABSTRACT

There has been a surge of Data Science programs offered by various business schools in North America and around the world in the last three years. Some of these programs are accessible in a classroom setting, whereas others are given online. Most of them cover basic concepts and algorithms and are offered in combination with software training. However, the cost of these programs has skyrocketed triggering confusion surrounding the value of these programs. A successful Data Science program must include extensive and challenging applications developed from real world business and other disciplines. The teaching instruments and pedagogy need to be adapted to the unique theoretical and applied requirements of the modern Data Science education. In the talk, I will discuss unique challenges facing Data Science teaching, as well as discuss ideas to create for students an effective classroom experience.

Our discussion is illustrated with examples from the Data Intelligence program at John Molson School of Business in Montreal since 2008. After years of engagement, it seems clear that the software industry such as SAS has at significant and constructive role to play in bridging universities and industry goals in order to create innovative educational programs.

INTRODUCTION

Following Wikipedia, “Data science, also known as data-driven science, is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, Darr (2013), Leak (2013), similar to data mining.” Statistics forms the foundation and is the predecessor of data science, hence, it is revealing to understand the differences between the two disciplines. Hand (1999) provides an early review of the similarities and differences between data science and statistics noting that while they share the common interest in discovering patterns in data structures, they go about how to achieve their goals differently. Shmueli (2010) gives further insights into the different paradigms of the two disciplines.

Watts (1991) suggested that the main reason behind students’ difficulties in learning statistics was its abstract nature. One cannot visualize random variables, probabilities or test a hypothesis simply by drawing, or by other illustrative methods. In contrast, in calculus, the derivative or integral can be explained by showing successive steps converging to the limit.

In data science, the learning obstacle can be overcome through data visualization and the interpretation of its real-life applications. In order to build a successful data science program, it is necessary to understand the differences between the modeling paradigms of data science and statistics. The sources of these differences are discussed in Section 1. Data science programs in North America are reviewed in Section 2. The positive effects of joint industry and academic support of the data science program at John Molson School of Business (JMSB) is discussed in Section 3.

SECTION 1: ARE DATA SCIENCE AND STATISTICS DIFFERENT?

The difference between data science and statistics can be explained, according to Shmueli (2010), by their distinguished focus on predictive and explanatory models, respectively, where the construct X is assumed to cause construct Y, via unknown function F, so Y = F(X). In statistics we look for function f such as that E(Y) = f (X). The purpose of explanatory modeling is to match f and F as closely as possible. The data (X,Y) provides the means for estimating f, and enables testing the hypotheses about f=F.

In contrast, in predictive modeling, the focus is on the prediction of Y, given X. In fact, even if the underlying causal relationship is indeed, Y=F(X), a function other than an estimate of f, may be preferable for more accurate prediction. Following Hastie et al (2009), the Expected Prediction Error (EPE), based on a quadratic loss function, can be decomposed as follows

$$\begin{aligned} &EPE = E\{Y - \hat{f}(x)\}^2 \\ &= E\{Y - f(x)\}^2 + [E(\hat{f}(x)) - f(x)]^2 + E\{\hat{f}(x) - E(\hat{f}(x))\}^2 \\ &= Var(Y) + Bias^2 + Var(\hat{f}(x)) \end{aligned}$$

The resulting EPE consists of the Variance of Y (first term), the bias (middle term) due to miss-specifying the model f and the variance (the last term) due to using a sample to estimate f. Shmueli (2010) elucidates that “The first term is the error that results even if the model is correctly specified and accurately estimated”.

The above decomposition reveals a source of the difference between explanatory and predictive modeling. In explanatory modeling, the focus is on minimizing bias in order to obtain the most accurate representation of the underlying theory.

In contrast, Shmueli states that “predictive modeling seeks to minimize the combination of bias and estimation variance, occasionally sacrificing theoretical accuracy for improved empirical precision”. This means that the “less correct” model with higher bias but lower variance may have lower EPE and could therefore be chosen from data mining perspective.

The discrepancies between explanatory and predictive paradigms exist in every step of the modeling process: data collection and preparation, exploratory data analysis, choice of variables and modeling approach, model selection and validation.

Lack of attention to the predictive and algorithmic approach to modeling among statisticians has been raised by Breiman (2001): “There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models.” As a result, there are still few courses and textbooks suitable for teaching introductory data science courses whereas there are many for statistics. Given the exploding interest in data science and its branch, artificial intelligence (AI), it is a time that they be combined or replace an outdated statistical curriculum that dominated our universities for a century.

Data science and statistics begin with the same object – the data – but the questions that they attempt to answer, the terminology and the approaches are different.

Table 1 summarizes the differences between the data science and statistical approaches to modeling.

Modeling steps	Data science	Statistics
1. Study design and data collection	• Observational design • Whole population is available • Large amount of data: examples and features	• Experimental design • Only sample is available • Small amount of data: observations and variables
2. Data Preparation, Exploration, Visualization and Dimension Reduction	• Imputation of missing values, studying missing value indicator • Outliers handling through transformation or binning, rarely removal • Cleansing and transforming variables • Data partitioning into training, validation and testing samples • Response definition is often done on an ad-hoc basis • Visualization methods are employed without hypothesized distributions • Dimension reduction like in PCA provide new variables for model fitting	• Imputation of missing values, no interest in the indicator of missing values • Outliers handling through transformation or removal • Cleansing and transforming variables • Rarely data partitioning is done due to scarcity of data • Response is defined through underlying theory • Visualization methods are to confirm assumed distributions • Dimension reduction like PCA is often used in confirmatory data analysis
3. Model development	• Models are developed based on accuracy of prediction • Multi-collinearity is not an issue for prediction • Models are prone to overfitting	• Models are developed to estimate and interpret the average effect based on all data • Multi-collinearity inflates error estimates
4. Model validation and selection	• Models' robustness or predictive accuracy is validated on independent samples. • Models are selected based on ROC or Lift curves	• Models are evaluated based on how well they fit the data, e.g. R2 in regression • Models are selected based on their fit and interpretability

Table 1. The differences between data science and statistics approach to modeling

Students' exposure to data science gives rise to questions such as:

- Why aren't we taking the sample from the population?
- Why aren't we testing the parameters?
- What is the role of the p-value in variables' selection when everything is significant.
- Should we be concerned with multi-collinearity in the data?
- How do we interpret variables in the selected model?
- How do we know the model is good?
- How do we know if it predicts well?
- How many good models are out there?

SECTION 2: LEARNING DATA SCIENCE

A 2011 McKinsey report estimated that up to 190,000 data analytics positions in the US will be created by 2018 and further 1.5 million more will be needed to use the findings for effective decision making. Canada's Big Data Consortium published a similar report in 2015. According to the Institute for Advanced Analytics, US universities created approximately 175 graduate programs, most of which were initiated in the last 3 years. These degree programs are categorized into two main groups: Business Analytics and Data Science. The Business Analytics programs offer interdisciplinary education in the areas of applied mathematics, statistics, computer science and business disciplines, like database marketing.

However, there are far fewer undergraduate data science programs compared to graduate ones in the US and Canada. A notable exception is the Data Intelligence minor program at John Molson School of Business of Concordia University in Montreal, Canada. It consists of the following four courses: data mining techniques, programming in SAS, linear models and forecasting. The enrollment in the first two courses skyrocketed by 500% and 300% over the last 3 years, respectively, see Figure 1.

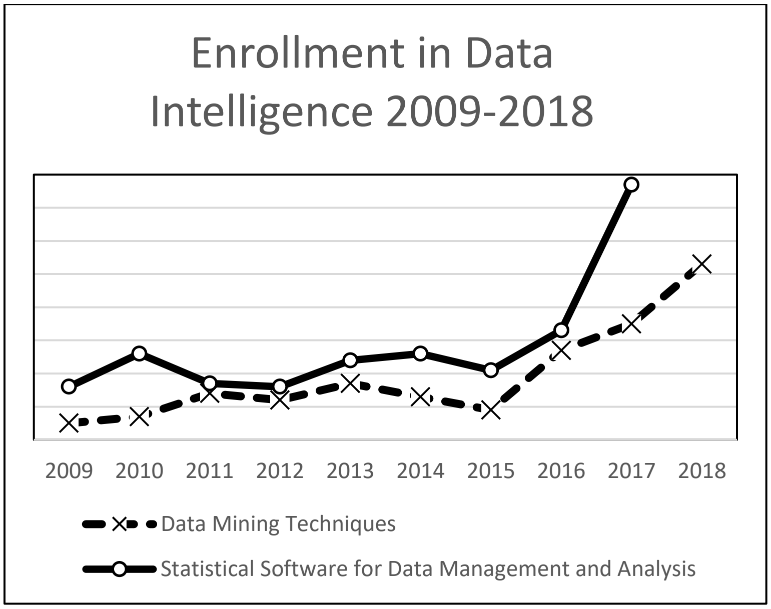

Figure 1. Enrollment in Data Intelligence courses has skyrocketed since 2016

SECTION 3: AN ACADEMIC-INDUSTRY VIEW ON LEARNING DATA SCIENCE

The data science goal of best prediction may help guide the development of appropriate learning curricula that are grounded in algorithmic foundations, programming and which are strongly connected to solving real-life problems.

In order to facilitate purposeful learning, students must work on prediction problems in the realistic context of big, messy data coming from fragmented and disconnected sources with ill-defined variables, outliers and missing values. The analytical problems may come with structured data or unstructured data such as text and image.

The presence of term projects in the curriculum provide an excellent training opportunity for students to work collaboratively and to familiarize themselves with the messy data of real life business problems. It is recommended that students search and find a suitable business problem and it associated large data from many data hubs available on the internet. The instructor's role is then to help guide the students' project definition so as to support a positive outcome.

Teamwork is promoted throughout the course, in lab exercises and in homework assignments. All classes take place in the lab using SAS®, SAS® Enterprise Guide and SAS® Enterprise Miner software. Active learning is stimulated by bonus points awarded for small tasks in the lab or as take-home activities.

The combined industry and academic experience of the Data Intelligence program is supported by the Joint JMSB-SAS® Certification in Business Analytics awarded to students of the Data Intelligence program. In addition, students may also seek Base and Advance SAS programming certificates from SAS®.

Furthermore, students' participation in the local data science conferences, poster sessions, and SAS meetings, such as Montreal SAS® User Group (MONSUG) is encouraged and is formally recognized within the data science course.

A new research initiative in data science sponsored by the Concordia University Part-Time Faculty (CUPFA) created and fulfilled a new industry-applied research assistant opportunity. Similarly, the demand for student teaching assistant positions and internships in data science with the local companies are now fast-growing.

As a result of creation of the multi-thread learning environment, students' interest in data science and self-learning initiatives have increased exponentially in the last three years. Students spontaneously organized the Data Intelligence Society of Concordia (DISC), a new organization that aims to promote data science in the university and the local business community. DISC sponsors SAS® software learning seminars, data modeling case competitions and local data talks. Remarkably, it is now professors who participate at the conferences organized by students.

The students' feedback from the program has been overwhelmingly positive which has translated itself into the enormous increase in enrollment.

CONCLUSION

Learning data science at the undergraduate or graduate level requires a thorough understanding of how different it is from traditional statistics. Focusing on predictive algorithms and on software programming in a real-life problem solving is a key to a successful development of data science program.

Described in some detail, the Data intelligence program at JMSB (Concordia University) is an example of the successful implementation of a data science curriculum. The program has been intertwined with actual business applications throughout the coursework, as well as has enabled students to actively participate in both academic and industry-related events.

SAS® Canada has played a very positive role in supporting the program through the use of its software, as well as through the creation of the much-coveted Joint Certification in Business Analytics. Finally, it has also contributed to the community by promoting local events and career opportunities.

The key to the success of data science programs is by expanding the breadth and depth of its curriculum and to create opportunities for students to understand and solve actual business questions with the state-of-the-art software tools like the ones offered by SAS®.

REFERENCES

Breiman, L. 2001 Statistical modeling: The two cultures. *Statistical Science*, Vol. 16, 199–215.
Canada's Big Data Consortium (2015). https://smith.queensu.ca/ConversionDocs/MMA/big-data-gap.pdf
Dhar, V. (2013). "*Data science and prediction*". Communications of the ACM. 56 (12): 64. doi:10.1145/2500499.

DISC, Data Intelligence Society of Concordia. http://disconcordia.com/

Hand D.J.,1999, *Statistics and Data science: Intersecting Disciplines*, *SIGKDD*, Vol 1, Issue 1.

Hastie et al., (2009), *The Elements of Statistical Learning Data science, Inference, and Prediction*. 2nd ed., Springer-Verlag.

Institute for Advanced Analytics, http://analytics.ncsu.edu/?page_id=4184

Leek, J. (2013-12-12). "*The key word in "Data Science" is not Data, it is Science"*. Simply Statistics.

McKinsey (2011). *Big data: The next frontier for innovation, competition, and productivity*. https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

Shmueli G., 2010, *To Explain or To Predict?* Statistical Science, Vol. 25, No. 3, 2010, 289–310
Shmueli, G. et al, 2018, *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R,* Wiley, 2010, Ed. 2nd, ISBN: 9781118879368
Watts D.G., 1991, *Why is Introductory Statistics Difficult to Learn? And What Can We Do to Make It Easier?* The American Statistician, 44(3) November 1991, 290-291.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Krzysztof Dzieciolowski
Concordia University
Krzysztof.Dzieciolowski@concordia.ca