Coffee With a Data Scientist: Tuhin Chattopadhyay, Ph.D.

In the fourth issue of DZone's Coffee With a Data Scientist, we had a chat with business analytics evangelist, Tuhin Chattopadhyay, to glean some of his expert insights and opinions on the Big Data space.

by Michael Tharrington · Aug. 04, 17 · AI Zone

In the fourth issue of DZone's *Coffee With a Data Scientist*, we had a chat with business analytics evangelist, Tuhin Chattopadhyay. Having had a wide range of experiences—from being a professor at multiple business schools, to working as the Editor-in-Chief at International Journal of Business Analytics and Intelligence, to offering business consultation as an Associate Director at Nielsen—Tuhin has viewed the business analytics world through a variety of different lenses.

A recent recipient of "Best Analytics and Insight Leader of the Year" at Big Data Analytics and Insights Summit 2017 and featured in India's Top 10 Data Scientists in 2016 by Analytics India Magazine, Tuhin kindly agreed to answer some questions for Coffee With a Data Scientist, offering up his expert insights and opinions on the space.

For those of you new to *Coffee With a Data Scientist*, our goal is to interview various data scientists and professionals in the field working on projects in machine learning, deep learning, data analytics, and/or big data in an effort to learn more about data science from the people who know it best. Oh yeah, and the coffee aspect of it all... we always like to offer our interviewees a coffee. So, if you're a data scientist who would like to share your thoughts on the subject and you'd enjoy a cuppa on us, please get in touch.

Tuhin manages to squeeze a couple of those big data fingers into the tiny handle of our specially crafted *Coffee With a Data Scientist* DZone mug in this expertly shot photo. Rock that mug, Tuhin!

***DZone*: You've had loads of experience in education with many years teaching as a professor in multiple B-Schools. What spurred you to move from the academic world to a more industry-focused position as a Data Scientist?**

Tuhin: After spending a decade in academics, you know by heart the different problems that you come across in books and how to go about solve them in a classroom situation. You also become completely cognizant of the statistical techniques that need to be applied to the problem. And, the questions that the students are going to ask are easily assumed after a decade of experience. So, there was not much challenge left in terms of solving an unknown problem in the class other than the research part of academics which I thoroughly enjoyed and poured my heart and soul into.

I developed a passion for the industry when I started taking on consultancy assignments related to analytics for businesses during my academic life. I really loved visiting these businesses and interacting with multiple stakeholders to better understand the business problem then subsequently translating the same to an analytics problem. Thus, my shift of gear from academics to industry is a kind of natural evolution via my passion for consultancy. I fell in love with the industry, as every business problem is a new challenge. The data is messy and mostly unstructured. Any of the hundreds of existing algorithms might be applicable or a new algorithm might need to be created to solve a unique problem. Thus, it was the challenge to journey on a completely unknown sea that brought me to a more industry-focused position.

Most of our readers are experienced developers but relatively new to the field of Data Science and Business Analytics. Being an academic, what are your suggestions for a path that experienced devs should follow to get ahead in this field? How can they best leverage their current dev knowledge, but use it to work toward a more data-science centric career?

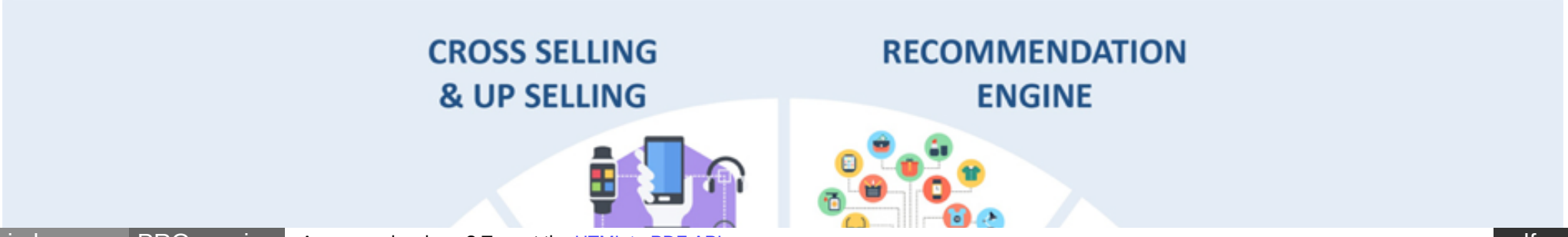
In any organization, analytics gain the momentum to fly from the strength achieved through both of its wings—big data technologies and machine learning techniques. Developers can easily equip themselves with big data technologies to kick-start a data-science centric career, but they must actively search for opportunities to leverage this tech in order to improve their data science knowledge and skills with machine learning algorithm. When specifically at progressing careers, there is a huge demand for big data architects who can create and manage the Hadoop ecosystem.

IMHO, experienced developers may initiate their expedition in the data science territory by developing with Distributed Filesystem (HDFS, Quantcast File System, Lustre file system, etc.), Distributed Programming (Flink released in March 2016, MapReduce, Pig, Spark, Storm, etc.), NoSQL Databases (HBase, Cassandra, Kudu, etc.), Document Data Model (MongoDB, ArangoDB, RethinkDB, etc.), Stream Data Model (EventStore, etc.), Key-Value Data Model (Redis, Time Series Database, etc.), Graph Data Model (Neo4j, TitanDB, etc.), NewSQL Databases (BayesDB, InfluxDB, TokuDB, etc.), SQL-On-Hadoop (Hive, Impala, MRQL, etc.), Data Ingestion (Kafka, Flume, Sqoop, etc.), Service Programming (Zookeeper,

Avro, Thrift, etc.), Scheduling & DR (Falcon, Oozie, etc.), Machine Learning (Mahout, Oryx, Deeplearning4j, etc.), Benchmarking And QA Tools (Apache Hadoop Benchmarking, Yetus, etc.), Security (Knox Gateway, Ranger, etc.), Metadata Management (Metascope, etc.), System Deployment (Ambari, HUE, Cloudbreak, etc.), Applications (Sphinx, PivotalR, etc.) and Development Frameworks (Cask Data Application Platform, Spring XD, Jumbune, etc.).

With years of experience as the Editor-in-Chief of International Journal of Business Analytics and Intelligence, you are a leader in Marketing and Customer Analytics. How do you see data science improving marketing and customer service in the coming years?

During my last five years' experience as Editor-in-Chief of IJBAI, I witnessed how both the academic and the corporate fraternity is constructing the domain knowledge of the subject brick by brick, building upon and expanding what we know. It's intriguing to observe the rapid pace of diffusion of data science and the success of any industry depends on the adoption of the same. Amazon Go on the corner of 7th Avenue at Seattle is a classic instance of the future of marketing and customer service. The infographic, provided below, exhibits the common applications of marketing and customer analytics in the retail sector.



Can you describe a specific Business Analytics/Data Science project that you worked on which was a success? What strategies did you follow and do you have any tips or insights that could be applied to other projects?

During my initial days in the analytics practice, I was deputed to a leading Southeast Asian bank where I spent a couple of months to accomplish a critical data science project. The first and foremost criterion in bringing success for the project required putting on a consultant's hat to understand the business environment better while keeping the statistician in me in the backseat—I especially must do so when the business environment is completely unknown to me. Capturing the right STEP (social, technological, economic, and political) variables will help to precisely define the business problem. The entire effort of model development will go for a toss if the problem is not defined correctly. The second important criterion is to master the art of obtaining the right data and sanitizing the same. Data is messy in most of the scenarios. The right understanding of the business coupled with the mastery of data cleaning techniques will help to brave the challenge. Last but not the least, it's a good practice to apply multiple modeling techniques or algorithms to a single business problem to address it from multiple perspectives. For instance, I used three techniques each for mapping the path of conversion and attribution modeling. The decision tree, neural network, and logistic regression were used for mapping the path of conversion, while Monte Carlo Markov Chain (MCMC), Shapley Value, and Survival Analysis were used to accomplish attribution modeling. Feel free to download some of my delivered projects from http://www.tuhinchattopadhyay.com/the-delivery-leader.html.

Now, on to my suggestions for the budding data scientists... The corporate world of analytics is getting more crowded and thereby becoming more competitive and ruthless. You have to

be a battle-hardened soldier to survive and excel. Incremental development in terms of learning a new language or a machine learning technique is a necessary but no more a sufficient condition. What makes you stand out from the crowd is your ability to convey complex mathematics in simple language to clients. To simplify it to the client, you need to first be able to simplify the same for yourself. Similarly, folks who have started their career as analytics consultants should pick up the model development skills to add depth to their conversations with the client. Secondly, to be a future leader of analytics, it is critical to develop and subsequently, refine an independent point of view about the work you are doing so that you may critically assess, review, argue, and appreciate the same the way you do about a recent movie or sports you watched. Independent thinking is an art. You need to practice thinking to be a thinker.

What subjective criteria would you generally suggest to use when evaluating the success of a machine learning model? Can you provide us an example of a successful machine learning model and tell us what you think made it successful?

The first thing from a subjective standpoint is to assess the model from the business perspective. The success of a machine learning algorithm lies in its ability to solve the business problem which does not necessarily come with a higher accuracy. Secondly, the success also depends on the deployment of the model. Rationalization plays a major role during the deployment phase to assess which predictions need to be used. Besides

rationalization, the governance and auditing of the scoring of new data play a major role in the successful deployment of the model. Thirdly, even if the model serves the business purpose well, sometimes there can be misperceptions among the client which blemishes the perceived success of the model. Perception management and de-escalations are critical at times for the perceived success of the model. Last but not the least; it is pragmatic to subjectively evaluate the significance of the insights developed by the model.

We hear a lot about the success of IoT and automation relying on Machine Learning to improve and solve problems in these domains. Where do the techniques for applying ML to IoT and Automation differ? Or, do they? Or, is there overlap like in a Venn diagram—some shared techniques, some not?

Traditionally, computer programs help to automate complex tasks and processes. Application of ML techniques help in automating the automation by software robots through Robotic Process Automation (RPA). Start-ups like WorkFusion leverage Artificial Intelligence (AI) in achieving cognitive automation. Of late, the revolution is all about the automation of automating the automation. This is achieved through Automated Machine Learning (AutoML/AML) tools like TPOT (Tree-based Pipeline Optimization Tool), Auto-Sklearn, Auto-Weka 2.0, DataRobot, MIJAR, PurePredictive, Xpanse Analytics, etc. where all the activities of a machine learning project like preprocessing the data, feature engineering & feature selecting, figuring out the applicable machine learning algorithms given the data and the objective, optimization of the hyperparameters, developing the plausible models,

selecting the best algorithm post applying all the pertinent algorithms on the data, presentation of the results, and finally the deployment are all automated. An interesting case on the application of AML would be the development of customers' lifetime value (LTV) by Airbnb.

ML and AI contribute to the success of IoT in multiple ways. Predicting when a machine is in need of maintenance saves millions of dollars. Stream Analytics and HDInsight may be used for transformation and analysis of SensorTag (6LoWPAN, ZigBee, etc.) data in the cloud. IBM Watson uses cognitive computing, a subset of AI, in answering questions posed in natural language over IoT. Application of ML algorithms to IoT and Automation overlap if the business objective demands so. For instance, Automated Speech Recognition (ASR) for IoT is achieved when Recurrent Neural Networks (RNN) work along with a Connectionist Temporal Classification (CTC) layer.

There are many "gold standards" that are well-known in different sectors for solving specific analytic problems in specific domains. These often existed before the phrase "Data Science and Business Analytics" was coined. For example, we have Conjoint Analysis, RFM Analysis, etc. which are now rebranded under the umbrella Marketing Analytics. How do these techniques from the pre-analytics era differ from this new Data Science and Business Analytics era, or have they just been re-skinned to fit into the new terminology?

The traditional statistical techniques, primarily belonging to marketing research, are very

different from the machine learning techniques employed in marketing analytics. In marketing research, hypothesis testing still plays a crucial role where the objective is to infer about the population from the collected sample data. However, in the modern trade, the entire population data, for instance, all the customers of a store, are tracked and can be obtained easily at the click of a mouse. Thus, all the statistical techniques, which were hitherto required to infer about population from the sample data, are becoming obsolete fast in the domain of marketing analytics.

In marketing analytics, both the techniques to analyze data and the infrastructure required to manage the big data are different from the traditional marketing research. As far as the techniques are concerned, on top of the traditional statistical techniques, machine learning techniques employ a feedback loop to train the model better to represent the reality like deep learning models. From the infrastructure perspective, the Hadoop clusters are set to manage the big data on which the machine learning algorithms are applied and finally the right infrastructure is required to deploy the machine learning model in the production environment.

Now, for a hot question! Is obtaining a Ph.D. necessary to become a successful Data Scientist?

To understand whether a Ph.D. is required to become a successful data scientist, one needs to understand the role a Ph.D. plays in the life of a research scholar on an average for 4-5 years. The qualities that a doctoral program is supposed to inculcate, including the

development of a research bent of mind, an intellectual curiosity to find the truth, lots of patience, and a desire to contribute to the world's body of knowledge are immensely helpful to be a successful data scientist. Besides the psychological aspects, spending years with the research process provides a lot of experience in anticipating and managing the potential challenges at every step of research. It’s not that without pursuing the degree such qualities can’t be acquired, or, by flipping the argument, all those who have pursued the degree have gained these qualities... but, the probability of possessing these qualities increases for those who go through the grinding process of a doctoral curriculum. And beyond that, a Ph.D. definitely doesn't hurt when job-hunting!

There are many online and offline boot camps on Data Science that guarantee an aspiring student can become a data scientist after the course. However, almost all of them have a similar course structure where they teach around 10-15 algorithms. Is this enough knowledge for a person to build a career in Data Science? How would you choose a practical Data Science course and do you have any suggested courses for DZone readers to look into?

Amazing question. First, the budding data scientists need to understand that learning a science is far more than learning a few algorithms. In fact, there cannot be a comparison. Rather, in the era of automation, there are systems which are smart enough to identify which algorithms are applicable given the data and the objective. Thus, the data scientists who only rely on 10-15 algorithms will have a very short professional career going as automation rules

them out. As far as my opinion is concerned, one has to have a strong foundation in research methodology, statistics, and computer science to build a sustaining career in data science.

Interesting experiments are going on across the globe with the course curriculum of data science. As business analytics do not work in silos, a harmonious blend of statistics, technology, and business management are required to do complete justice to the breadth and depth of the subject. Carnegie Mellon University's Heinz College offers Master of Information Systems Management: Business Intelligence & Data Analytics (MISM-BIDA) that allocates 72 units for Analytics and Technology Courses , 42 units for management courses and 18 units for capstone project. In India, Post Graduate Diploma in Business Analytics (PGDBA) is a 2-year full-time residential program with four semesters in Analytics and Data Science jointly offered by three premier institutes of the country – Indian Institute of Management (IIM) Calcutta, Indian Institute of Technology (IIT), Kharagpur, and Indian Statistical Institute (ISI), Kolkata. The students will spend the 1st semester of six months at ISI Kolkata to learn statistics and machine learning theories for analytics. The 2nd semester will take place at IIT Kharagpur to engage with the technological aspects of analytics. The 3rd semester will be at IIM Calcutta to focus on the application of analytics in functional areas of management. Finally, the students will be required to do an internship of six months duration on an analytics project in a business organization. Following the same philosophy, MIT Sloan School of Management offers Master of Business Analytics joining hands with the MIT Operations Research Center. Thus a student of business analytics should be open to immersing themselves in multiple disciplines to help them develop a holistic perspective.

You have worked with many organizations to better their understanding of how analytics can benefit the business. What standards or processes do you follow when setting up a Business Analytics Centre of Excellence and a great team to make the implementations successful at an outside organization.

Leonardo da Vinci drew the helicopter in 1493, 450 years before the actual helicopter would take to the air. The ability to foresee the future, having an innovative mind, and being a researcher at heart are all critical traits for the success of an analytics centre of excellence, keeping all other technical skills above the acceptable level.

On the other hand, the implementation team at the client organization requires a completely different set of credentials. Folks should be open to embrace a different culture, should be excellent in communication and coordination, able to always display a positive attitude, be friendly and dissolve even the slightest traces of ego, again assuming relevant technical skills above the acceptable level. These people are the face of their organization at the client's level. So, they will be responsible enough to present the right image about their organization to the client. Finally, they need to delight the customers and one of the strategies they may follow is to under-promise and over-deliver.

Had you ever come across DZone previously? As an expert data scientist, what are your suggestions for improving our coverage of Machine Learning and Data Science to meet the needs of data professionals?

Yes, I do follow DZone regularly and have deep regards of the breadth and the depth of the topics that you cover. Here are my two cents to improve it further. You may conduct a focus interview/ panel discussion on a pertinent topic in analytics. That would capture the varied perspectives that would otherwise be mono-dimensional in a 1-on-1 interview. An extension of this suggestion is to organize conferences in each of the technologies you cover. You may also come out with physical magazines with more enriched content than what is available online.

Is there anything I haven't asked you about that you'd like to add? (Cool or interesting happenings in Machine Learning that you want to mention? Shout-outs to others in the field you'd like to recognize? etc.)

All is not well with AI and analytics. There are some growing concerns with both AI and the inputs and the outputs of analytics. Even keeping the recent heavily-publicized exchange between Mark Zuckerberg and Elon Musk aside, there's no doubt that like any other technology, AI also must be innovated with caution to drive it in the right direction. As far as the input of analytics is concerned: the data of where you are going, how much time you are spending, with whom you are spending your time, what you are eating, what soap you use in your bathroom—every tiny detail of your personal life is tracked! Thus, privacy is highly compromised in the era of analytics. As an output of analytics, the marketing is customized to manipulate the minds of customers.

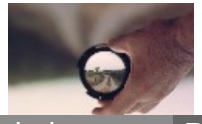
Analytics leveraged for the social sector should gain more traction in controlling crime,

managing traffic, conserving the environment, etc. More organizations like Bayes Impact, SocialCops, Outline India, and DeepMind should come forward to make analytics work for the social good. More conferences like Do Good Data from Possibilities to Responsibilities, Artificial Intelligence for Social Good, AI for Good Global Summit, and AI for Social Good should be organized to address the burning problems like poverty, hunger, health, education, and the environment. The data science curriculum should offer a specialization in social innovation analytics. Many more AI and cognitive computing competitions should be organized like the $5 million IBM Watson AI XPRIZE. Foundations like the AI for Good Foundation should come forward to channelize the power of analytics to make the world a better place to live in.

Thanks for the interview, Tuhin.

If you missed the last issue of *Coffee With a Data Scientist* with Lillian Pierson, check it out!

Like This Article? Read More From DZone



Data Science Start-Ups in Focus: H2O.ai

What Is a Data Science Workbench and Why Do Data Scientists Need

One?

What Are Data Scientists, and Are They Here to Stay?

Free DZone Refcard

Recommendations Using Redis

DOWNLOAD

Topics: AI , BIG DATA , DATA SCIENCE , INTERVIEW

Like (4) Comment (0) Save Tweet

AI Partner Resources

[Article] What Is Artificial Intelligence for IT Operations (AIOps)?
BMC

Six Ways to Boost App Engagement With AI
Neura

Read the Gartner Market Guide for AIOps Platforms
BMC

New skill sets for Aritficial Intelligence-powered IT Ops

BMC

Maximizing Energy and Cost Savings with AI

Neura

Improving Digital Health Tools with AI

Neura