What does a Data Scientist do?

I asked some former colleagues and they provided me with the following:

Answer #1

1 Data Scientists work with customers to derive insight from the customers' data

2 Data Scientists use many tools. Everyone has their favorite language and data manipulation tool. But, a data scientist easily adapts to new tools. Often a data scientist must use the tools that the customer wants.

3 A data scientist often has a "hacker" mentality. This "hacker" mentality implies that the data scientist will get a job done quickly. Given the data-centric view of the data science task, the methods are usually not easily reusable. A data scientist does not try to create robust reusable solutions for many different data sets. A data scientist wants to get to the heart of a given data set or small set of data sets.

4 Data scientists often work with other BI experts like DBAs and Analysts and software programmers.

4a A Data Scientist will ask a DBA to help find data in a database, shape the data, and formulate queries against a database. Many data scientists like doing this work themselves.

4b A Data scientist will ask an analyst what kind of insights are desired and how should these insights be presented. Often a data scientist must educate an analyst as to what kind of insights are possible from a given data set.

4c A Data scientist will ask a programmer (software developer) to create software that productizes a process for gaining insight into a data set. The programmer will make the Data scientist's process robust and reusable.

5 A data scientist must work closely with the owner of the data. The data scientist will ask questions to get more information about the data. The data scientist will listen to customers to understand what the customer hopes to get out of the data.

Answer #2

Researchers at a famous east-coast pharmaceutical company want to automate the detection of cancerous blood cells. They take pictures of the blood smears under a microscope. One picture can be many Giga-pixels in size. The blood was stained with multiple stains that stain different cellular features (cell membrane,

nuclear membrane, nuclear DNA, etc.). The images are segmented and processed. For each cell, numerous attributes are recorded. The patient's diagnosis (Blood Cancer/No Cancer) is recorded along with these attributes. The data set was handed over to a data scientist who then proceeded to identify the best algorithm that could predict if a patient had cancer based on the image of the blood smear. The data scientist had to try out many algorithms with a variety of parameters. The data scientists tried to normalize differences in image capture between images.

Answer #3

Business Case: the data scientist is modeling healthcare data to help hospitals predict readmissions of patients, and predict development of new diseases/complications.

The hospital gathers lots of data – lab tests, prior history, demographic information, it also utilizes studies performed on certain diseases (e.g. patients with diabetes are more likely to have certain accompanying complications).

The data scientist's ultimate goal is to build 2 models – one is going to calculate a risk score for a patient to be readmitted, and the other one – to calculate risk score/probability of a patient developing a complication. The data scientist tries various algorithms (independently) and discovers that decision tree algorithm is well-suited for readmissions, while logistic regression gives the best performance for complications.

Now, data scientist needs to 'explain' results to nurses, showing some indicators as to why this particular patient has a high risk of complication and/or readmission. Let's consider the Systolic blood pressure variable. It is treated as a continuous one (patients can have values of 110, 120, 121 etc.). Both algorithms (though used the same input data) independently bin continuous variable in the way that suits each algorithm better. Note, that bins have the same labels: 'low', 'normal', 'high' and 'very high', though the split points are different. Then, the patient with blood pressure X can be placed into 'normal' bin for readmission model and 'very high' bin for complications. Very confusing for nurses!

Possible solution – bin such 'potentially-dangerous' variables into clinically meaningful categories before running algorithms on top of them. This way it is ensured that nurses would see bins labeled as 'hypertension', 'pre-hypertension', 'normal', 'hypotension' in their system. And the same patient won't have different Systolic BP label in different models!

Extra-Curricular Data Science

Here are some links to organizations that support data science meet-ups. A meet-up is usually a local 1-3 hour meeting at no cost:
http://datasciencedojo.com/community/meet-ups/
http://www.meetup.com/data-science-dojo/
http://tdwi.org/Home.aspx
http://www.tdwi.eu/home/
http://www.networkworld.com/article/3037521/big-data-business-intelligence/career-boost-break-into-data-science.html