Coursera: Final Assignment

Course: What is Data Science?

Author:
Raghuvar Prajapati
Data Scientist & Full-Stack
Developer

October 8, 2018

Contents

1 Introduction/Executive Summary

Data Science has always been a fascinating aspect for me since the day I heard the buzz around about it. So, just after graduating I started looking for data scientist position desperately. And well I get into one which now is driving me crazy and kept me excited all the times with their findings, insights and the information that came out of it. While working as a Data Scientist I came across that let's start with the very beginning of it and find out what were the grounds which made it so-called ***"the sexiest job of the 21st century"***. And I came across this course on coursera which leads me to take this course and it's been an amazing experience while learning the very basics of it. I have completed this entire course in just 2-3hrs. I'll look forward to taking all the related courses of the same in upcoming days.
This report includes the very basics of what is data science? what are the scopes of it? what you need to have at least to get started with this buzz and many more such thing.

2 Data Science: The sexiest job in the 21st Century

Yeah, you heard it right. Data science is becoming one of the sexiest jobs for this generation and upcoming generation. Its popularity and demand are increasing exponentially if I'm not wrong. And finding a true data scientist is becoming rare because people are exaggerating this term too much and they are not looking at the basics or inside into it. It's like I'm using it but I don't know what am I using. Almost over the past few years, the demand for data scientist has become necessary for almost all kind of organization. Because the number of data whether it's structured, unstructured or semi-structured has been increased rapidly. A report by Mckinsey Global Institute says there's going to be a shortage of skilled data analyst of around 140,000 - 190,000 alone in the United States. So, shortage of data scientist makes companies look for the other organization which can solve their problems related to data, like Kaggle, kdnuggets, DrivenData, CrowdAI, H2O.ai, KNIME and some other startups which are enthusiastically solving the problem in the data domain.

3 Standard to be a Data Scientist

The term itself is so fantasizing that people often get perplexity with it. Although there are many definitions to it as a data scientist is someone who finds the solutions to problems by analyzing big or small data using appropriate tools and then tells stories to communicate their findings to the relevant stakeholders. As one of the famous book author Dr. Vincet Granville defines data scientist as one who can easily process a 50-million row dataset in a couple of hours and who distrusts models. As he clearly put the statistician and data scientist into two different boxes. So, as per my understanding if one has following standard he/she can be a data scientist:

- A curious mind. Eager/Excite about the data.
- Fluency in Analytics.
- Ability to communicate the findings whatever it is.

4 Regression

Regression is one of the methods to find the relationship between the variables. One which is changing for different instances in a process with other makes an impact on the other variable.
Example: Book a cab ride: Base fare would have been decided anyway. So, how much distance you travel and how much time you spent(if you get stuck in let's say traffic) then paid money. So, time-money and distance-money are the parameters which are going to increase in the base fare.

5 Data Mining

It involves some sets of steps as:

1. **Establishing Data mining goals:** The cost-benefit trade-offs for the desired level of accuracy is are important considerations of data mining goals.
2. **Selecting Data:** Type of data, its size and frequency of collection have a large impact on the cost.
3. **Preprocessing Data:** Identify the irrelevant attributes of data and expunge(something unwanted or unpleasant) such attributes from further consideration. Deal with missing data also whether it's missing systematically or randomly.

4. **Transforming Data:** Reduce the number of attributes needed to explain the phenomena by transforming data from one to another.e.g. PCA can reduce the no of attributes without losing the information.
5. **Storing Data:** Choosing of storing the data scheme such that it should facilitate efficiently reading from and writing to the database. Data safety and data privacy are the primary concern for storing the data.
6. **Mining Data:** After appropriately processed, transformed and stored now is the timing for data mining. It includes data analysis methods including parametric and non-parametric methods, machine learning algorithms. A good starting point is data visualization.
7. **Evaluating Mining Results:** Formal evaluation of the results extracted from previous steps includes: testing the predictive capabilities of the models on observed data to see how effective and efficient the algorithms have been in reproducing the data. Known as **"In-Sample Forecast"**

6 Deep Learning

Examples: Speech recognition, face recognition (Multiple level of neural network)

- **For neural networks:** linear algebra is must
- **For Data Scientist:** Probability and statistics, databases, physics, computer science, math, How to program, computational, algebra and calculus, different statistical distribution, relational databases.
- **Applications of Machine Learning:** Predictive analysis, recommemdation system, classification, regression, precision and recall

7 Data Science in Business

1. Start capturing the data, Data Archiving.
2. A data scientist should have Curiosity, storytelling, and sense of humor then technical skills. Ability to communicate it with a story(present your findings).
3. A brilliant data scientist who is passionate about the field of IT won't necessarily excel in the field of healthcare if they are passionate about it.

4. Applications of Data Science: Relief program is done by developing countries to send over the foods and other necessities by analyzing tons of data. Basically optimizing resource allocation for relief aid. Determining if an application for a credit card should be accepted based on the applicant's financial history and data.