Pre-Course Study Pack: Data Science

Day 1: Tuesday, 26th November 2019

Welcome to Stutern Graduate Accelerator Data Science Course.

Lesson 1: Thought leader's review

As you embark on the journey to become a data scientist - the sexiest profession of the 21st century - a good starting point will be to gain some understanding on what some thought leaders think about the profession. Therefore, I have put together a collection of articles to help wet your appetite:

- Article 1: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century
- Article 2: https://www.bernardmarr.com/img/bigdata-case-studybook_final.pdf
- Article 3: https://flowingdata.com/2009/06/04/rise-of-the-data-scientist/
- Article 4: https://www.oreilly.com/radar/drivetrain-approach-data-products/
- Article 5: http://www.datascienceassn.org/sites/default/files/Building%20Data%20Science%20Teams.pdf
- Video: https://www.youtube.com/watch?v=RHsO10q7e2Y

You will come to realise that the structure of every data science project follows a similar pattern that includes - data collection, data cleaning, data transformation, model training and application deployment. Time spent on each phase of "the Data Science Process" totally depends on the motivation behind the project. For some projects, you find that the most important task is to engineer a data collection system. While for others, it might be more important to spend considerable time in training a model that gives highest accuracy in prediction.

Also, I am a strong believer of the saying "Begin with the end in mind". Therefore, I would like for you to start thinking about the motivation towards your final project. For example, I had my undergraduate background in Electrical Electronics Engineering before I stumbled on the Data Science profession. For this reason, my first application of the data science process was to look at electrical energy consumption data for an electric distribution company in the UK. I was able to build an energy disaggregation model that takes whole house electricity consumption data from a smart meter to detect the consumption of each appliance within the house.

TASK:

IN NOT MORE THAN TWO HUNDRED AND FIFTY (250) WORDS DEVELOP A MOTIVATION REPORT TOWARDS YOUR FINAL COURSE PROJECT.

Note: The goal is not to think about what tools or resources for the project but to build an intrinsic base to help make the project a success for you.

Day 2: Thursday, 28th November 2019

Let's gain first intuition as a Data Scientist

Lesson 1: Exploratory Data Analysis (EDA)

The foundation of data science is in basic statistics and probability theorem. During the course, we will spend some time to look at both aspects of mathematics and even further learn as well as build complex mathematical models on these concepts.

But to get you going, I would like to introduce an elementary approach that you will always refer to in your data science process - Exploratory Data Analysis (EDA). EDA was introduced by John Turkey whom is often referred to as the God Father of data analytics. He defined EDA as

"an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe that are there."

In other words, it is an approach for you to understand the problem you are trying to solve or not solve by diving straight into the data. The basis of which come from application of summary and graphical statistics to data that helps to form hypothesis for further analysis towards knowledge creation.

Additionally, EDA helps to develop a relationship with your data. It helps to gain intuition, understand the shape of it, and try to connect your understanding of the process that generated the data to the data itself.

Here is a resources to help you understand some basic practices of EDA

https://www.dropbox.com/s/q2qkcyvtu6s5oak/chapter4.pdf?dl=0

You would have noticed that the on-boarding test given to you was to perform an exploratory data analysis on a given dataset. While some of you did pretty good job, I believe there is always room for improvement.

TASK:

LOOK AT OPTIMISING YOUR ON-BOARDING TASKS. USE ANY TOOL THAT IS CONVENIENT FOR YOU (R, PYTHON, SPSS, POWER BI, EXCEL, ETC) AND CARRYOUT THE FOLLOWING ACTIVITIES:

1. For each variable, create a summary statistics that tells you about the data type
2. For each variable, use a summary metric that helps you describe the data
3. For each variable. provide a graphical representation of the data distribution
4. Create a new variable "average_score" represented from "math", "reading" and "writing" scores
5. Create another variable "average_score_cat" that categorises the "average_score" using WAEC grading system
6. Find a relationship (if any) between each variable and the new variable "average_score_cat"
7. Create graphical representation of the relationship(s) discovered in 6.
8. Develop an hypothesis about which variables that can help predict the "average_score_cat" of a new student.

West Africa Examination Council (WAEC) Grading System

Grade	Min score	Max score
A1	85	100
B2	70	84
B3	65	69
C4	60	64
C5	55	59
C6	50	54
D7	45	49
E8	40	44
F9	0	39