

|  |  |
| --- | --- |
|  | **UNIVERSITI TUNKU ABDUL RAHMAN** |
|  | **Assignment 1** |
| Course Code: | UECS3213 / UECS3453 |
| Course Name: | Data Mining |
| Lecturer: | Dr. Simon Lau Boung Yew |
| Academic Session: | 2019/01 |
| Title: | Industry Talk |

|  |  |  |  |
| --- | --- | --- | --- |
| Name | I.D. No | Course | Practical Group |
| Tan Ying Yao | 1703648 | SE | P1 |
|  |  |  | Mark: /10 |

**Marking Scheme**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No. |  | **Poor** | **Adequate** | **Proficient** | **Subtotal** |
|  |  |  |  |  |  |
| 1 | Language clarity and | not | not | reasonable, |  |
|  | formatting | reasonable, | reasonable, | concise, |  |
|  |  | not | concise, | understanda |  |
|  |  | concise, | understanda | ble |  |
|  |  | not | ble |  |  |
|  |  | understand |  |  |  |
|  |  | able |  |  |  |
|  |  | 0 - 10 | 11 - 20 | 21 - 30 |  |
| 2 | Comprehensiveness / | incomplete, | partially | complete, |  |
|  | Completeness of analysis | lack of | complete, | informative |  |
|  |  | information | with some |  |  |
|  |  |  | information |  |  |
|  |  | 0 - 10 | 11 - 20 | 21 - 30 |  |
| 3 | Critical analysis | illogical, | partially | logical, |  |
|  |  | incorrect | logical, | critical, |  |
|  |  | assumption | correct | correct |  |
|  |  |  | assumption | assumption |  |
|  |  | 0 - 10 | 11 - 30 | 31 - 40 |  |
|  |  |  |  | Total | /100 |
|  |  |  |  | Assessment | /10% |

**Summary:**

Data science is a multi-disciplinary field that uses scientific methods to extract knowledge from data similar to data mining. Dr. Chan Zan Kai, a machine learning engineer at Axiata Digital Advertising presented his talk regarding the application of data science in advertising. Dr. Chan explained that internet advertising had overtook its competitor in advertising in recent years. Digital advertising is the act of marketers collecting personal data to create relevant advertising to the consumers. Traditional advertising usually relies upon ad agency to relay their advertisement to the audience but digital advertising specialise in a specific group of consumers. These are usually dependant on the advertising ID provided by companies such as google which tracks each person’s data.

Data can be classified into three category – First party data, second party data and third-party data. First party data are usually high-quality data in small scale which is attributed to each client. Second party data are bought from first party and are pre-processed. Third party data are collected in large scale and provides terrible quality. These data sources are usually from Facebook and Google partners and filtered through ad exchange platform and finally shaped into proper advertising for the consumers.

Finally, audience are then segmented based on demographic, situational, psychographic and geographic location. These categories help provide relevant and useful advertising to their potential target consumer. There are also places where ICO credit scores are used to classified consumers. These scores dictate the percentage of a user’s activity and categorise them as such to provide advertising to them.

**The latest trend in the industry:**

The year 2019 marks the beginning of business across the world begin the meshing of the physical and virtual mode. The field of Data Science is transitioning with advanced data technologies taking over business processes for better efficiency and productivity. The role of a “data scientist” will be far more diverse and expand into a greater scope. There will be an increase in business automation but this does not mean that data scientist will be obsolete as these tools are also beneficial to the human experts. Automated Machine Learning systems will take predictive analytics to greater levels but is very complex to learn. Chatbots, Virtual Reality (VR), and Augmented Reality (AR) will revolutionize product and service marketing. These AI-enabled tech will personalise customer experience through live demos, interactive simulations and accessible visualization of solutions. Blockchain are now mainstream and covers diverse industry sectors from banking and finance to health insurance. These blockchain are important in data protection and fraud detection and ensures good data management.

The title of data scientist is also fluctuating as a result due to the larger scope of and roles involved. These causes the term of “data scientist” being more limiting as result and cannot be used to explained the role of the job. Regarding the job-market trends in data science, it can be inferred as a double-edged sword. The advancement I technologies will erase the need for data scientist yet these machines are only useable by said data scientist. The field is still evolving and in a state of flux resulting a change being made frequently. Digital marketing, an offshoot of Data Science has also revealed that the market will eventually rely heavily on AI to understand customer behaviour. According to Lisa Matherly, VP of Content Marketing at McAfee, this will create more opportunities for personalized customer experiences. AR and VR will also revolutionise marketing and customer care as more businesses such as IKEA’S VR Kitchen providing live demos, simulations and virtual service for the customers. Many businesses had also implement and experiment with Chatbots, VR, AR, and automated processes for improved efficiency and increased customer satisfaction. Videos had account for 80 percent of internet traffic and thus can be utilised in digital advertising.

When asked about the trend of the current industry in Malaysia, Dr. Chan Zan Kai mentioned that companies had stopped hiring for these data scientist. He also mentioned that it has slowly become a niche field as it has a high potential of returning high profit. The field heavily relies on the data acquired, but these data are only valuable to certain companies that thrives on it.

**Skills needed to be a data scientist:**

Data scientist requires a wide array of technical skills which are required to collect the data necessary for their operation. Thus, it is recommended for data scientist to have programming skills, specifically in python. The knowledge of programming is vital in deciphering and organise a large set of data. The knowledge of analytical tools are also valuable insights to siphon and analyse data obtained. SAS, Hadoop, Spark, Hive, Pig, and R are some of the most popular data analysis tool that data scientist frequently use. Certification are required to establish expertise in the use of these analytical tools. The job also required one to be adept at working through unstructured data. A good data scientist should be able to understand and manage data that is coming unstructured from different channels.

A good data scientist should also possess some non-technical skills that can help aid their data collecting process. For example, a strong business mindset will help them to acquire the know-how of the elements that make up a successful business model. The power to discern the problems and potential challenges that need solving for business to sustain and grow are also crucial. They should also have great intuition. This can be explained as being able to perceive patterns where none are observable on the surface and knowing the presence of where the value lies in the unexplored pile of data bits. This greatly increase the efficiency of their work experience and productivity.

When Dr. Chan Zan Kai were asked with this question, he responded that a good data scientist should possess a business mindset. To clarify he suggest that the companies nowadays mainly hire business school graduate as they are familiar with the business background. Dr. Chan also reprimanded the importance of good programming skills such as being fluent in the programming language of python. He also explained that being a data scientist is not that rare and anyone can become one if he puts in the work.

Questions asked during the talk:

1. Are data collected personal or grouped together?

A: Data collected are grouped together, the more common interest people share, more data are received. The location is known so tracking down each individual is very easy. These data are sold as money is priority in data collecting.

1. Do (Virtual Private Network) VPN help hide one’s identity?

A: We may track wrong data but there will still be data collected. Sometimes there are even data collected from North Korea.

1. How do you collect these personal data?

A: The information required are sent into the DNS and from then on the location of data is known.

1. Do we have ICO in Malaysia?

A: No, not yet anyways.

1. Could you explain what is the meaning of GBM?

A: It is called a Gradient Boosting Machine. Through trial and error we obtain the data required.

1. What is the job scope for a data scientist? Is feature extraction a requirement?

A: The job is mainly collecting data and processing data. Feature extracting is not common and thus not considered.

1. How would you rate your job currently?

A: Data scientist is a profit-oriented job. All data collecting is with profit in mind and that is how the transaction works.

**Your critical analysis / opinion of the presentation/technology/etc:**

Data science is an ever-growing industry with no clear definition defining it’s job scope. The field is constantly transitioning and expanding as advanced data technologies starts taking over the business processes for better efficiency and productivity. The role of data scientist now encompasses a huge array of know-how and technical skill. Machine Learning systems are still highly dependant on skilled data scientist to fully utilise. The rise of Virtual Reality (VR) and Augmented Reality (AR) also helps the industry to grow greatly. The advancement of technologies constantly changes the definition of a data scientist making it an ever-growing and fluctuating industry.

The presentation of Dr. Chan Zan Kai brings to light that the fact is internet advertising is slowly dominating the market share while TV growth slows down. Digital advertisement are mainly to collect personal data and to track users. He also explained that data are grouped together when obtained from the demographic such as sex, age and occupation. Data can also be classified into three category – First party data, second party data and third-party data. First party data are usually high in quality but collected in small scale. Second party data are bought from first party and are pre-processed before available for sale. Third party data are collected in large scale but are usually terrible quality-wise. Dr. Chan clarifies that the you are tracked through advertising ID such as smartphone or through web surfing. The data collected are segmented through categories to personalise data advertising.

The technology of data science is still growing and improve continuously. Digital marketing, an offshoot of Data Science has established that the market will eventually rely heavily on AI to comprehend customer behaviour. This AI implementation will create more opportunities for personalized customer experiences. AR and VR will also revolutionise marketing and customer care as more businesses such as IKEA’S VR Kitchen providing live demos, simulations and virtual service for the customers. Many businesses had also implement and experiment with Chatbots, VR, AR, and automated processes for improved efficiency and increased customer satisfaction. Videos had account for 80 percent of internet traffic and thus can be utilised in digital advertising.

In conclusion, Data Science is still a thriving industry and more fine-tuning are still required to better define its role in it’s industry. It is here to stay as the future will be heavily reliant in data mining. Digital Advertising is merely the logical offspring of data science as it heavily depends on the data and information it receive.