Faculty of Information Technology
Summer Semester B, 2020
FIT5145 Introduction to Data Science
Assignment 3 – Thinking Critically about Data Science

Student Name : Gayatri Aniruddha
Student ID : 30945305

Context 1: Understanding the professional data scientist

Question:
Do you believe the Australian market for data scientists has become inconsistent or unstable?
Why or why not? (8 marks)

Answer:
Data science is in demand in the entire world and Australia is no different. Data Science is the hottest career to get into right now and this trend is only going to grow exponentially with the advent of newer applications and developments of newer tools to aid the field. Australia has a good, stable and consistent market for data science. The Australian market is exploding with jobs thanks to the emergence and development of machine learning, big data and artificial intelligence. (Pash, 2018)

Data is used in all aspects of our everyday life and in all the industries from banking, retail, media, entertainment, transportation and recruitment. There is no dearth of data and we need data scientists to collect, wrap, analyze and present the data for our convenience and future predictions.

Data scientists have opportunities in every field and the biggest opportunity is in the field of finance. There are plenty of startup options too. Alternatively, data scientists can work for a university and perform cutting edge research. Similarly, data scientists can also use their skills to work for government bodies for creating new advancements in data science. The minimum starting salary of a data scientist with no experience is around 50K annually. The companies are now recognizing the importance of harnessing this ever-increasing wealth of information.This figure goes further increases with experience and on average a data scientist with 5 years of work experience makes a minimum of 150K annually. (Sinclair-Jones, 2018)

Question:
What is your opinion on whether Australia has a data science skills gap? (5 marks)

Answer:
In my opinion, Australia does not have a data science skills gap. There are numerous programs available in top universities like University of Melbourne, University of

Sydney, University of Queensland and Monash University in the field of data science both at the bachelors and masters level. (Ray, 2017)

The Masters program in data science equips the students with the knowledge of programming, algorithms, modelling, statistics and natural language processing. Students even have the opportunity to indulge in research and take part in industrial internships. The best part is that this program is open to students from non IT background as well which helps them to use their existing knowledge from their field and learn the skills needed to perform the job of a data scientist. (Brown, 2009)

These universities not only offer high-quality education but have connections with the industry and academia which help the students secure a good placement after completing their education.

Question:
Choose a specific case study of a real-life data science project and explain to them your rationale for your choices. You will need to:
• Explain the case and the business value sought.
• Provide definitions of these roles and their differences and explain your reasoning behind why you have chosen those definitions.
• Explain how the data scientists would interact with other professionals, and why data scientists are even necessary.
(12 marks)

Answer:
Here, we consider the example of Freebase which is a huge knowledge-base which contains different types of machine readable data. The data present here is organised and structured. It is collected from numerous online sources and many individual wikipedia contributors. The main reason for the development of this website is to provide users around the globe with some common resource that allows them to access information in a more effective and systematic manner. It was developed by Metaweb, an American software company and has been made available publicly since March 2007 and was acquired by Google in 2010. The business value from this project is that the textual data has been utilized for developing projects like Named Entity Recognition and Natural Language Question Answering. (Wray, 2018)

Here, Data analysts and data engineers worked in close connection with data scientists. Data analysts are people who develop insights with data. They understand statistical theory and use that to provide practical solutions to industry problems. Data engineers are people who are involved in data infrastructure and automation of data processing. They maintain data architecture. In the example of freebase, data analysts were the ones who were involved in collecting data. They researched the different online websites and individual contributors from whom they could collect relevant information. The data engineers were involved in building systems that could store and process this large amount of data collected. Their role was to integrate any new data management technique into existing structures.

Finally, data scientists are people who develop data models. They interact with data analysts to extract useful and meaningful insights from data. They interact with data engineers to develop data centric projects. In freebase, they were involved in collaborating the different data collection methodologies with data analysis methodologies. They worked in close connection with their owners and stakeholders to improve the functioning of freebase. We need data scientists in all stages of collecting, wrapping, analyzing and presenting data. Data scientists have sufficient understanding of statistical theory, machine learning, programming, and can talk to other professionals to understand and visualise situations. Thus, the demand of data scientists is very high and they are necessary in all the industries.

Context 2: The business of big data

Question:
Do you think that the value Uber provides to people more generally justifies some of these criticisms? (8 marks)

Answer:
I agree that the value Uber provides to people does justify some of the criticisms against Uber. Customers are able to travel faster and cheaper when compared to normal taxis. Public transportation does not run round the lock and not all taxi companies run 24 hours a day. People travelling late at night are able to get a safe ride back home, thanks to uber. At the same time, riders have the option of sharing their rides with other riders going in the same direction through UberPool. Uber previously experienced some incidents of drivers attacking the riders. Hence, Uber now conducts proper driving and criminal checks for all it's drivers. Uber conducts checks that date back to upto seven years. (Lisa, 2019)

People living in the city who do not need a car on a daily basis can save money by using Uber. They need rides for weekend outings and other occasions. With the advent of uber eats, it's now becoming easier for people to get their food delivered at their doorstep, practically making it easier for them to live without a car. Uber also helps its drivers by providing them flexibility. Drivers can now earn on their own schedule. Uber has even begun to provide its drivers in the European Union with health care, accident insurance, maternity and paternity leave. Thus, Uber is one of the easiest applications to use and is of great help to both it's riders and drivers.

Question:
What type of analytics do you think Uber's dynamic pricing model uses? (2 marks)

Answer:
When a lot of people in the same area request for rides at the same time, the prices may be higher than normal to attract more drivers to the area. Uber's dynamic pricing algorithm uses predictive analysis and revolves around a number of factors which

include, time and distance of the route, traffic conditions and the current demand for drivers.

<u>Question:</u>
How do you think that data would likely move through the standard value chain? What data is this? (5 marks)

<u>Answer:</u>
Uber collects (COLLECTION) the data through our login and ride details. It has access to our name, contact number, email address, credit card details, home address and work address. Uber's data collection system is constantly and dynamically updated and Uber states that the data is anonymized and aggregated. (Bajpai, 2020)

Uber stores (ENGINEERING) our information using built in data encryption so that the data can't be stolen from our account. Uber knows where it's customers are waiting and where it's drivers are stationed (GOVERNANCE) and ensures that they are put together fast. (WRANGLING)

Once Uber analyzes (ANALYSIS) it's supply and demand, it implements it's dynamic pricing algorithm to decide it's surge pricing and boosts fares at peak times to attract more drivers to the riders. When a rider requests for a ride, it presents (PRESENTATION) all possible options i.e UberX, UberXL, UberPool, UberSUV, UberBlack etc to gain results and value. (OPERATIONALISATION)

<u>Question:</u>
Discuss why data management and wrangling are a critical aspect of Uber's participation in the sharing economy. You will need to explain the difference between curation and wrangling and refer to at least two curation models. (10 marks).

<u>Answer:</u>
Data wrangling is the process of transforming raw data into a form which is valuable for analytics. Here, we gather, select and transform data to answer various analytical questions. In the case of Uber, data wrangling involves collecting the different rider information, driver details, location of different riders and drivers. Uber then converts this raw data into a more systematic format in order to allocate the different drivers to riders and decide the pricing. This data wrangling is essential because working with raw data is challenging.

Data management is the management of data. It involves developing the value of the data. It ensures that data is trusted, reusable and easily accessible. In Uber, data management revolves around managing the rider and driver profiles. It ensures that the rider details are easily available to the driver and vice versa. There is increased productivity as the riders are able to book a ride from anywhere at any point of the day. There is no data duplication and this leads to cost efficiency. With appropriate use of data management uber can respond easily to market changes and fellow

competitors. Uber's data management system ensures that our information is secure and our information is backed up.

In data curation, data is collected from diverse sources and integrated in order to enhance it's value. Here, data is managed throughout its lifecycle, from its creation, storage, archival to its destruction. Main aim of curation is data reuse and retrieval. Data curation lifecycle model mentions the activities required to curate data throughout its lifecycle. This digital data curation model revolves around creating, accessing, using, selecting, disposing, ingesting, preserving, reappraising, storing, reusing and transforming the data. Another data curation model is the DataONE lifecycle. It revolves around planning, collecting, assuring, describing, preserving, discovering, integrating and analysing the data.

References

Zhou, Z. H., Chawla, N. V., Jin, Y., & Williams, G. J. (2014). Big data opportunities and challenges: Discussions from data analytics perspectives [discussion forum]. *IEEE Computational Intelligence magazine*, *9*(4), 62-74.

Sinclair-Jones, Julia.(2018, July 26). *Data Science in Australia: Overview, Salary Data and Career Path.* Retrieved from https://www.themartec.com/insidelook/data-science-in-australia

Pash, Chris.(2018, November 2). *5 amazing Australian jobs available right now in big data.* Retrieved from https://www.businessinsider.com.au/5-amazing-australian-jobs-available-right-now-in-big-data-2018-11

Brown, G. (2009). Review of Education in Mathematics, Data Science and Quantitative Disciplines: Report to the Group of Eight Universities. *Group of Eight (NJ1)*.

Ray, Tanmoy. (2017, December 29). *Best Masters Programs in Data Science & Analytics in Australia | Top Australian Universities for Big Data Analytics.* Retrieved from https://www.stoodnt.com/blog/best-masters-programs-in-data-science-analytics-in-australia-top-australian-universities-for-big-data-analytics/

Kim, M., Zimmermann, T., DeLine, R., & Begel, A. (2016, May). The emerging role of data scientists on software development teams. In *2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE)* (pp. 96-107). IEEE.

Buntine, Wray. (2018, January 8). *Case Studies of Data and Standards.* Retrieved from https://www.alexandriarepository.org/syllabus/introduction-to-data-science/64776/

Cohen, P., Hahn, R., Hall, J., Levitt, S., & Metcalfe, R. (2016). *Using big data to estimate consumer surplus: The case of uber* (No. w22627). National Bureau of Economic Research.

Goetz, Lisa. (2019, November 17). *4 Reasons Why Riders Choose Uber.* Retrieved from https://www.investopedia.com/articles/markets/063016/4-reasons-why-riders-choose-uber.asp

Martin, Nicole. (2019, March 30). *Uber Charges More If They Think You're Willing To Pay More.* Retrieved from https://www.forbes.com/sites/nicolemartin1/2019/03/30/uber-charges-more-if-they-think-youre-willing-to-pay-more/#296716527365

McGuffog, T., & Wadsley, N. (1999). The general principles of value chain management. *Supply Chain Management: An International Journal*.

Bajpai, Prableen. (2020, January 12). *How Uber Uses Your Ride Data.* Retrieved from https://www.investopedia.com/articles/investing/030916/how-uber-uses-its-data-bank.asp

Orchard, S., Kerrien, S., Abbani, S., Aranda, B., Bhate, J., Bidwell, S., ... & Chatr-Aryamontri, A. (2012). Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nature methods*, *9*(4), 345-350.

Ryan, G. W., & Bernard, H. R. (2000). Data management and analysis methods.

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., Van Ham, F., Riche, N. H., ... & Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, *10*(4), 271-288.