

**ASSIGNMENT FRONT SHEET**

**Course Name: ALY 6000 Introduction to Analytics**

**Professor Name: Dr. Prabagaran Santhanakrishnan**

**Student Name: Dong Quoc Tuong (Lukas)**

**Student Class: Fall 2019 CPS Term: A. 2019**

|  |
| --- |
| **Module 5: Case Study Leveraging Big Data**  **Completion Date: October 21 st Word Count: 1486 Due Time:12:00am** |

**Statement of Authorship**

*I confirm that this work is my own. Additionally, I confirm that no part of this coursework, except where clearly quoted and referenced, has been copied from material belonging to any other person e.g. from a book, handout, another student. I am aware that it is a breach of Northeastern University’s regulations to copy the work of another without clear acknowledgement and that attempting to do so renders me liable to disciplinary procedures. To this effect, I have uploaded my work onto Turnitin and have ensured that I have made any relevant corrections to my work prior to submission.*

**Tick here** to confirm that your paper version is identical to the version submitted through Turniti

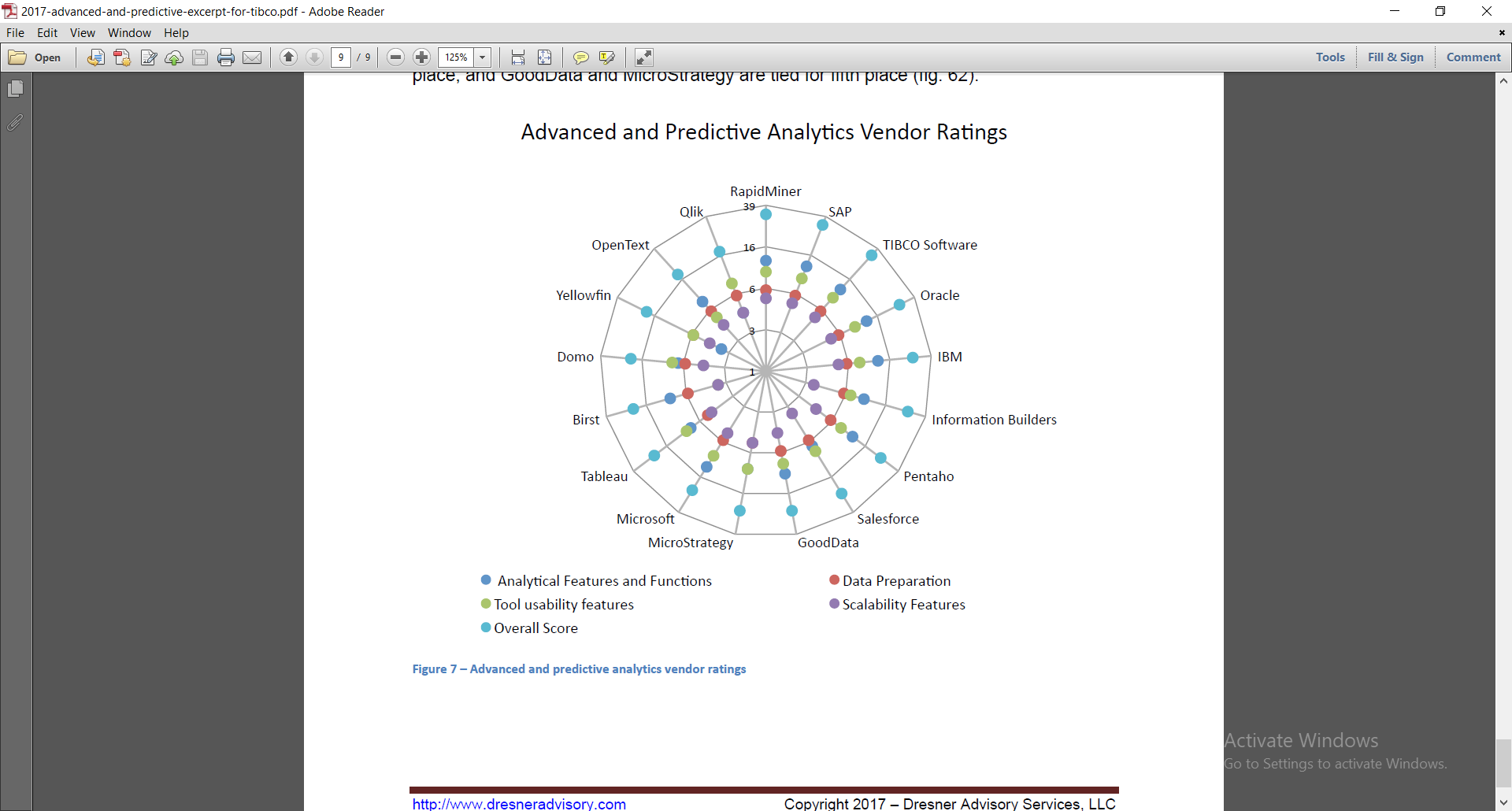
1. **Introduction**

Gartner Research predicted that data volume will increase by 8 times over the next 5 years but only 20% of them are structured. (Attaran & Attaran, 2018) At the turn of the century, data-driven industries have experienced tremendous growth thanks to the massive explosion of video and photo data, astonishing smartphones usage’s the amount and the exponentially connected world through the Internet. By the year 2020, approximately 1.7 megabytes of new data will be established every single second for every individual across the world, thus rending the global data pool to more than 44 trillion gigabytes. Unsurprisingly, more and more companies are changing their business models to seize such a lucrative influx of information about their potential customers. One of the ways to that company does to sense of the sea of information is through “Predictive Modeling”.

1. **Analysis**

“Predictive modeling” can be defined as the collection of mathematical techniques employed in seeking mathematical relationships between a target, response, various predictors, “dependent” or “independent” variables to evaluate those predictors ‘accuracy if they are embedded into these relationships in the future. Due to these relationships’ imperfect nature, in reality, it is always good to give some measure of uncertainty for the predictions, particularly those with an assigned confidence level of more than 95%. Another task that users have to take care of is “model building”. Normally, predictor variables can be classified into three groups: “Not affecting respondents”, “Affecting respondents”, which are definably included in the test, and “May affect the respondents under the right conditions”. (Dickey, 2012) The last group is the trickiest one and methods to test whether to include those variables or not have been in development for years. It is not wrong to say that there is much to learn about “Predictive modeling” in the field of data analytics. It is the way in which big data, the epitome of the 4th industrial revolution, drives the decision-making process for smart business operations. There are a wide variety of ways that “Predictive modeling” can collect customer’s data:

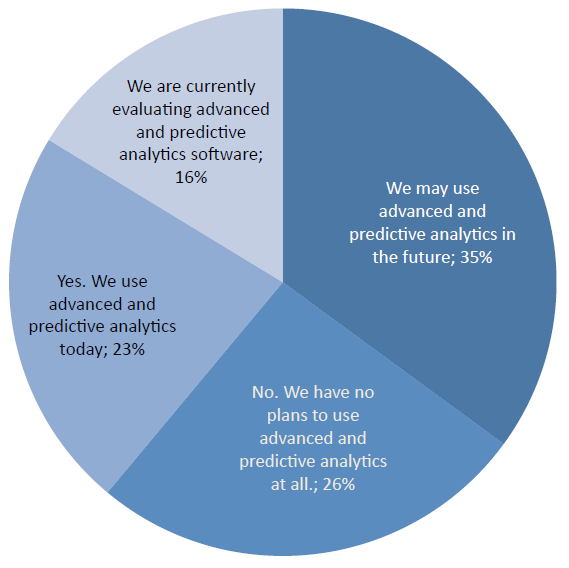
For the customers, tracking technologies have enabled many companies to shift from the dominated aggregate data analyses to individual-level data analysis. Their main goal is to boost the number of rows that assigned with transactions per customer and monetary values. The more successful they are in connecting the transaction or monetary values to customers, the more likely they will complete a customer’s profile. Such process is further amplified by the widespread loyalty programs. A web that connects everything together will start to emerge, for example: customer transaction data with financial background, onsite visitation with purchases’ amount, etc. If one decides to share social media’s information like Facebook or Instagram then the customer-level database becomes extremely nuanced and abundant for exploitation. (Edwards, 2019). Equally important, the increasingly data-rich business environment has created two trends in recent years in products manufacturing. First, the reduction of unpopular or unwanted products. Second, the widening of the row width (product information matrix) to accommodate all the necessary or available information about the products. Retailers can have the more dynamic, descriptive or even micro-targeted products towards the buyers. Sometimes, they can even gather enough information about a new product even before introducing it to the public



**Rating of common data analytics tools (Advanced and predictive analytics market study, 2017)**

What about the time? Unlike the data aggregated to monthly, weekly or even annually level, the contemporary retailing business comes with a timestamp which enables a continuous assessment of customer behavior, product assortment, stocking information, etc. For instance, if a shopkeeper wants to know how their discount or product reallocation has performed so far. A simple database that demonstrates the connection between the in-store guest’s movement and their purchases can address these concerns effectively. Furthermore. The increase of touchpoints (online shopping channels) allows brands to gather more information about their customers. Retailers are starting to explore new territories like Showrooming (offline showroom and online buying) and Webrooming (customers showing the opposite behaviors than predicted). There have been discussions about implementing an empirical model to control the customer’s conversions to different channels through various common touchpoints. (Bradlow, Gangwar, Kopalle, & Voleti, 2017). Lastly, data about the spatial location of the customers at any given time improves the effectiveness of any marketing campaigns substantially by allowing providers what they can offer and how they can offer it best. If the geo-spatial location is linked with the CRM database, retailers can have access to the guests’ purchasing history and can identify the profitable location to open their stores. However, these short-term revenue-maximizing schemes might nurture some unethical drawbacks that might make the customers reconsider about the potential with the hyper-localization of a product or a brand.

One of the best illustrations of how data analytics and data science benefits both the consumers and the businesses at the same time then look no further than Nexflix. Being one of the largest television networks in the world, it is natural for Nexflix to store millions and millions of gigabytes of information on each individual user, 53 million across 50 countries to be exact since the company was founded in 1997. AI technology has proven to be one of the most fundamental tools for Nexflix as it is estimated to save $1 billion annually for the company. Furthermore, the use of CDN allows Netflix to reduce bandwidth cost and the content backup to Google cloud storage enables the company to recover any data that was mishandled within hours. Going upon and beyond just simply recommending films base on customers’ preferences, Netflix also produces its own TV shows that are beloved much critics and audiences alike, such as “House Of Cards”, “Orange Is The New Black”. This is nothing short of a miracle that the prediction algorism is accurate and inspired other streaming providers like Hulu, Amazon to do the same. Netflix created an independent stream of revenue that is arguably more competitive than the traditional TV cable channels with the constant flow of feedback, helping them to further improve their original series’ content. (Nicolaus Henke et al., 2016)

  
**Survey on the current implication of predictive analysis in business setting. (Advanced and predictive analytics market study ( excerpt ), 2017)**

Nevertheless, there are a lot of hurdles that a company must overcome in order to recreate the aforementioned Nextflix example. First of all, the lack of good data makes it hard to connect employ “Predictive Modeling”. Business executives rarely know how to create a single customer data warehouse with unique IDs on everyone that include their purchasing patterns or demographic information, collected through various touchpoints. Secondly, in spite of its popularity as the primary tool for predictive analytics, the Regression model realizes heavily on the Assumption factors, a deadly sin. Every model has a degree of assumption but Regression is infamous for it. There are cases that these assumptions turned out to be completely false and caused severe collateral damages. Additionally, customers might change their purchasing patterns as time flies by. The once effective predictors can render useless. Therefore, it is essential to keep updating the old models was created years ago. Thirdly, data scientists might forget to include changeable elements in their predictions. One instance of this was the financial crisis of 2008 and 2009. Invalid models prediction mistakenly assumed how likely mortgage customers will pay without taking into consideration the non-stop rising house prices, causing an economic catastrophe on a global scale. (Davenport, 2014)

Those are some of the technical dilemmas that a business might encounter, but what about the ethical and privacy issues when data become so important nowadays. Some concern about the usage of data to purposely harm their financial wellbeing. Orbitz ( a hotel booking website) has been criticized for showing more expensive hotel deals to customers who access their website by Apple products instead of traditional PCs. Others worry that their voice recognition devices can pick up private conversations and store it to the cloud services Nonetheless, the biggest legal loophole is the lack of regulations on the reselling of data between aggregators/ providers. A clear answer to this is the opt-out option. Although in October 2016, the Federal Communications Commission passed the privacy legislation that forces the business to receive consent from their customers first before collecting data on web browsing, app usage, location, and email content, there are much left to be desired. (Bradlow et al., 2017) . Besides governmental intervention: businesses can also employ these two methods to convince their customers about their effort in respecting customers’ privacy:

* Showing the advantage of predictive modeling to their buyers (Transparency approach)
* Rewarding loyalty in exchange for data collection

1. **Conclusion**

To sum up, data prediction is the new frontier of business and whoever masters them, master the world. As a result, we must be very cautious in using data to not only benefit the human kinds, but also strengthen our freedom and democracy.

**References**

*Advanced and predictive analytics market study ( excerpt )*. (2017). Retrieved from https://www.tibco.com/sites/tibco/files/resources/2017-advanced-and-predictive-excerpt-for-tibco.pdf

Attaran, M., & Attaran, S. (2018). Opportunities and challenges of implementing predictive analytics for competitive advantage. *International Journal of Business Intelligence Research*, *9*(2), 1–26. https://doi.org/10.4018/IJBIR.2018070101

Bradlow, E. T., Gangwar, M., Kopalle, P., & Voleti, S. (2017). The role of big data and predictive analytics in retailing. *Journal of Retailing*, *93*(1), 79–95. https://doi.org/10.1016/j.jretai.2016.12.004

Davenport, T. H. (2014). A predictive analytics primer. Retrieved from https://hbr.org/2014/09/a-predictive-analytics-primer

Dickey, D. A. (2012). Introduction to predictive modeling with examples. *SAS Global Forum 2012*, *337*, 1–14.

Edwards, J. (2019). What is predictive analytics? Transforming data into future insights. Retrieved from https://www.cio.com/article/3273114/what-is-predictive-analytics-transforming-data-into-future-insights.html

Nicolaus Henke, Jacques Bughin, Michael Chui, James Manyika, Tamim Saleh, Bill Wiseman, & Sethupathy, G. (2016). *The age of analytics: compteting in data-driven world.* Retrieved from www.mckinsey.com/mgi.