**Digitalization and Big Data in Various Industries**

Laurence T. Burden

Purdue Global University

IN402: Modeling and Predictive Analysis

Steward Huang

December 18, 2023

**Review of “A Survey on Big Data Analytics in Health Care”**

Big data is a term that can be hard to definitively define. The exact criteria for information to transition from small or normal data to big data is not clearly established. Some characteristics are defined to let us know when we see a very large amount of data that it can be reasonably called “big data.” Still, big data has a large role to play in modern health care.

Seven characteristics can be used to classify big data: data source, volume, velocity, data types (variety), availability, architecture, and tools (Singh & Kumari, 2017). Each of these characteristics is different for big data versus normal data. First, we will look at the source of the data.

Traditional data consists of normal information that can be saved in a database. On the other side, big data sources are disparate and numerous. This could include spatial data, text messages, and audio/video content. All of these sources can be compiled together to form a big data repository.

The most obvious characteristic to define for big data is volume. This is the size of the data. Normal data sizes can go up to the lower range of terabytes for large databases. Big data will have numerous terabytes or even petabytes of information. This volume is possible due to the velocity of the data.

Velocity refers to the speed at which the data is created (Middelburg, 2023). Traditional data is created in a manner that can be handled in batches. Big data is constantly being generated in real-time.

The type of data refers to its structure. Traditional data is generally structured and can be stored in SQL databases. Big data may be structured, unstructured, or semi-structured. Big data is considered to always be available, whereas normal data may be confined to working hours. The architecture of normal data tends to be centralized, versus the decentralized nature of big data.

The tools used to process different levels of data also differ. Relational databases store traditional information. Big data requires a different mindset to parse through. Purpose-built data storage systems such as NoSQL and Hadoop are necessary.

The three Vs of high-volume, high-velocity, and high-variety are included in the characteristics just discussed. Different organizations may use a different number of Vs in their definitions. IBM describes 4 Vs and some other companies use 5 Vs (Singh & Kumari, 2017). Someone could describe big data with only two Vs. Volume and variety could be argued to be enough to describe a dataset large enough to be considered big data.

Modern health care relies on predictive analytics in several ways. The diagnosis of some diseases is assisted by analytics. The spread of a pandemic can also be modeled using predictive analytics. Some areas can still benefit more, though.

One area that could benefit is the creation of a recommendation system (Singh & Kumari, 2017). This would allow someone to choose the best physician based on their needs. Certain challenges still exist, though.

All big data endeavors encounter certain roadblocks and challenges to overcome. Health care is no different. Data preparation can be difficult, choosing efficient algorithms is a struggle, privacy concerns must be addressed, and selecting the right talent is vital to success (Singh & Kumari, 2017).

**Big Data Mining in Banking**

The world of data is exploding at an extreme pace. One estimate puts the amount of data created daily at 120 zettabytes (Duarte, 2023). Many industries can benefit from sifting through relevant information for key insights. Banking is among these industries. Many opportunities exist for the banking industry by implementing predictive analytics. Many techniques exist to execute this process.

Hassani et. al. describes five areas that can see value creation by using data mining within the banking industry: Security and fraud detection, risk management, customer relationship management, branching strategies, and efficiency evaluation (2018). A closer look at each of these will give insights as to how they can support growth in the industry. The first area looked at will be security and fraud detection.

Cybercrime is on the rise. This has put many online companies at risk of having their customer information stolen. Banks are no different, but they may have a bigger reason than most to protect their customers’ information. Having one’s banking details stolen means that the thief has access to your money. They must go to extraordinary lengths to ensure that this information is protected to keep others’ money safe.

Security and fraud detection machine learning algorithms would benefit from classification and K-mean clustering techniques. This would help the banks identify “…phishing, fraud, money laundering…” (Hassani, et. al., 2018), etc. The number of transactions made each day is too numerous to have a human review them. Using computers allows us to catch more criminals.

A bank takes on quite a bit of risk when they are entrusted with others’ money. They also have a responsibility to invest the money wisely to ensure it continues to grow. For this reason, banks can use the large amounts of investment information they have gathered to create risk management and investment strategies based on predictive analytics. One area that banks can place more effort in is an investment section called peer-to-peer lending (Hassani, et. al., 2018).

Risk management and investment banking also benefit from classification and k-mean clustering techniques. This allows the bank to quickly get a feel for the amount of risk involved in an investment opportunity. This could include analyzing a company or a person requesting a loan.

The bank can use information on their customers to better understand them and know what areas to improve on or to suggest other services to existing customers. “…80% of financial service organizations globally list customer experience as its top priority” (Hassani, et. al., 2018). Any company must retain and build its customer base to grow and make money. This sector also benefits from classification and K-means clustering.

The other sectors that can benefit from predictive analytics include branching strategies and efficiency evaluation. Branching refers to the individual branches of a larger bank. Many metrics can be used to measure and increase efficiency. Like the other sectors, classification and K-means clustering are useful techniques to help build models.

Banking is not immune to the need to incorporate predictive analytics. There are at least five sectors that can benefit from implementing classification and K-means clustering techniques. This will help the bank continue to grow and serve its customers efficiently.

**Review of “Role of Data Mining in Insurance Industry”**

Data mining is the process of sifting through huge datasets to determine patterns useful to the business’s use cases (Stedman & Hughes, 2021). Many industries can benefit from combing through the data they have available to make the company more efficient. The insurance industry is no different.

There are many areas of opportunity for data mining in the insurance industry. These include risk identification, customer analysis, marketing, reinsurance, and detecting fraud (Umamaheswari & Janakiraman, 2014). The first to be discussed is risk analysis.

Insurance companies are always looking to reduce risk. This ensures that they are profitable by finding insurance opportunities that are unlikely to need a payout. Data mining can help create models that identify the level of risk with any customer.

Customers expect to get useful services that are relevant to their needs. Insurance companies can implement a customer-level analysis to find these opportunities. These findings can then be used for marketing and reinsurance purposes.

Marketing needs to be tailored to best entice new members to join a service. Data mining can be used to create a model that finds the best characteristics of a marketing campaign. This will lead to increased profits for the company.

Fraud is an area of intense concern for companies that deal with financial concerns. An unscrupulous customer could financially benefit by convincing the insurance company to pay on a policy that legally should not have been activated. For this reason, data mining can be immensely beneficial to find these activities and flag them before the company can be harmed. The models can include patterns of past fraudulent claims and be used to predict the probability of a customer performing a new fraudulent claim.

Another opportunity that was not discussed in the paper is that of efficiency evaluation. The company can use the metrics captured within the company to determine how efficient various offices are at performing their duties. This will help the company know where to expend resources to bring low-performing offices up.

The insurance industry is not immune to the challenges associated with big data analysis. These challenges include noisy data, high volume, high complexity, corrupted data, and missing data (Umamaheswari & Janakiraman, 2014). Each of these challenges takes a unique process to overcome.

Data accuracy and standardization are vital to a useful model being created. Noisy data must be wrangled into a consistent data format. Any incorrect values must be identified and fixed. Finally, missing data must be dealt with. This can be done by dropping the rows with missing data or by imputing the missing data.

Plenty of opportunities exist for the insurance industry to utilize data mining and analytics. There are challenges that must be overcome first, though. Companies that can successfully navigate these challenges will be rewarded with more customers and higher profits.

**References**

Singh, V., & Kumari, M. (2017). A Survey on Big Data Analytics in Health Care. *Proceedings of the 11th INDIACom*.

Middelburg, J.-W. (2023, August 17). *The four V’s of big data*. Enterprise Big Data Framework©. https://www.bigdataframework.org/the-four-vs-of-big-data/#:~:text=Velocity%20refers%20to%20the%20speed,Variety%20of%20Big%20Data

Duarte, F. (2023, December 13). *Amount of data created daily (2024)*. Exploding Topics. <https://explodingtopics.com/blog/data-generated-per-day>

Hassani, H., Huang, X., & Silva, E. (2018). Digitalisation and big data mining in banking. *Big Data and Cognitive Computing*, *2*(3), 18. <https://doi.org/10.3390/bdcc2030018>

Stedman, C., & Hughes, A. (2021, September 7). *What is data mining?*. Business Analytics. <https://www.techtarget.com/searchbusinessanalytics/definition/data-mining>

Umamaheswari, K., & Janakiraman, S. (2014). Role of Data mining in Insurance Industry. *Compusoft*, *III*(VI), 961–966.