# INFS 4020 – Big Data Concepts (SP2 2022)

| | |
|---|---|
| Course | Big Data Concepts (INFS 4020) |
| Name | Gitae Bae |
| Email | baegy002@mymail.unisa.edu.au |
| Identifier | 110310861 |
| Date | Thursday, 09/06/2022 |

## Assignment 2 – Big Data Strategy Proposal

[Gitae Bae]

[09/06/2022]

# Contents

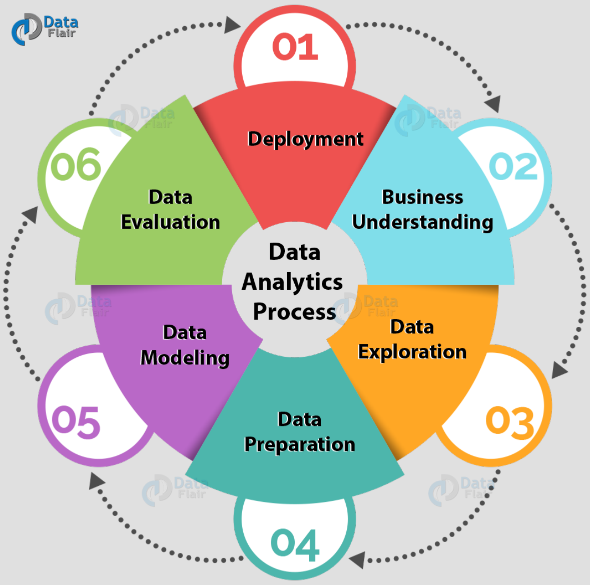
# Executive Summary

### INTRODUCTION

Recently, the concept of big data has attracted global attention as it has been applied and used in various social and economic fields. A structural model in which each dimension of a bank affects customer satisfaction, cross-sell intention, and relationship continuity intention is proposed.

### KEY PRIORITY CONSIDERATIONS

In this study, the effect of each dimension of bank service quality on customer satisfaction, relationship continuation intention, and cross-purchase intention is studied.



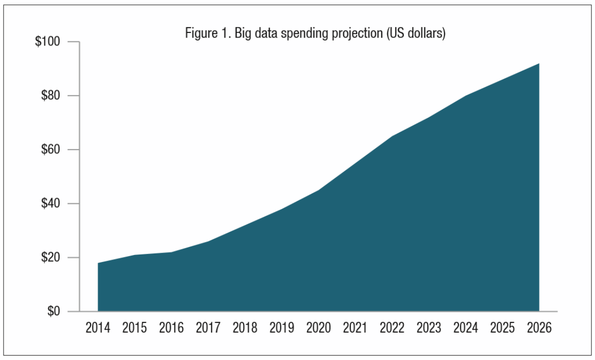
### DATA SOURCES

At the beginning of the project, data to be collected were selected through consultation agreement between data scientists and marketers.

### BIG DATA ARCHITECTURE

In the data extraction, processing, and purification stages, previous data analysis and big data analysis are typically differentiated.

**Soaring spending on data analytics: Spending on data analytics will grow at a compound annual rate of more than 14%, fuelling rapid advances in targeted marketing promotions, enhanced risk management, and fraud prevention**



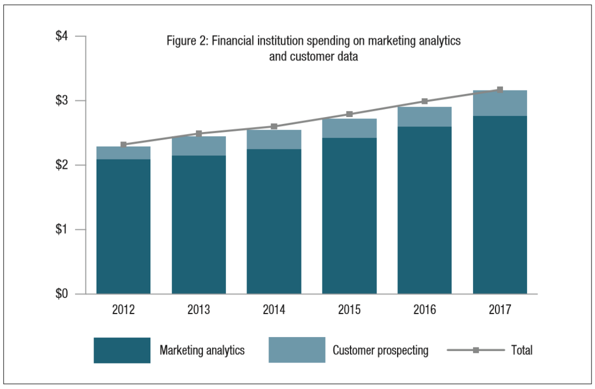
Source: Wikibon

The above data is available in full to RBWG members. For access, go to http://retailbanking.theasianbanker.com.

### BENEFITS AND CHANLLENGES

On the premise that the higher the customer perceives the quality of service provided by the bank, the more positive behavior is formed, the main success factor of the bank service is derived for customers using the bank

**Financial institutions are spending more on data analytics and using it for marketing**



Source: Aite Group

The above data is available in full to RBWG members. For access, go to http://retailbanking.theasianbanker.com.

### CONCLUSION

In future studies, research on the appropriate scale development and refinement process for measuring factors that are more detailed and reflect customer satisfaction should be continuously conducted.

Gitae Bae baegy002(110310861)

## Introduction

Recently, the concept of big data has attracted global attention as it has been applied and used in various social and economic fields. Big data is one of the important technologies that open up new possibilities for international development at the World Economic Forum (WEF) in 2012 and is core as a keyword for the IT industry worldwide. Companies and organizations are expected to be able to increase the timeliness and effectiveness of decision making through the appropriate use of big data and to develop productivity by improving work capabilities and stabilizing processes. In particular, it is expected to be able to be derived that business innovation, including the development of customized products and services, through active integration and utilization of big data. It was defined as a large volume, high velocity, and variety of information assets that required an information processing method, and was defined as 3V. Then, what about the use of big data in the financial industry? which includes banks, credit card companies, and insurance companies that we commonly encounter in our daily economic life? Commonwealth Bank is one of the most famous banks in Australia and will incorporate cross-selling into the bank and improve service quality to improve customer relations and attract more customers. Hence, here are some examples of key priority considerations, data sources, big data architecture, benefits and challenges of Commonwealth Bank of Australia. In general, the financial industry is not only used by the majority of the people, but also the amount of data inflow, possession, and aggregation is vast and the rate of increase is fast due to factors such as many types of products. Accordingly, it is expected that the range of big data utilization will be diverse and the value of use will be higher than in any other industry.

## Key priority considerations

The target company of this study is a case study of a big data-based target marketing proof of concept task conducted by Commonwealth Bank operating in Australia. Commonwealth bank's existing target marketing method was implemented as a rule-based expert system using traditional variables such as customer's age group, occupation group, and membership. This method aims to select target customers for marketing according to the marketer's intuition when executing short-term campaigns. Although the previous method was simple and could select target customers quickly, the results of the campaign showed limitations that did not exceed a certain level due to the limitation of not being able to provide services tailored to customer needs. Commonwealth bank tried to select target customers for marketing who reflected financial needs and situations based on customer behaviour by implementing target marketing based on big data analysis. The big data analysis in this case study is to generate a model that predicts dependent variables from explanatory variables generated from big data, using data from the past for several customers of commonwealth customers. The big data analysis environment was stored, processed, and analyzed big data with a commercial big data analysis platform based on Postgre-SQL and Map/Reduce, and text analysis was performed with analysis and classification with Python-based commercial text analysis tools. Big data analysis identifies customers' ability to subscribe to financial products, preferences for financial institutions and products, and relationships between financial institutions and customers.

Finally, target customers who can cross-sell financial products are selected and actual application is demonstrated.

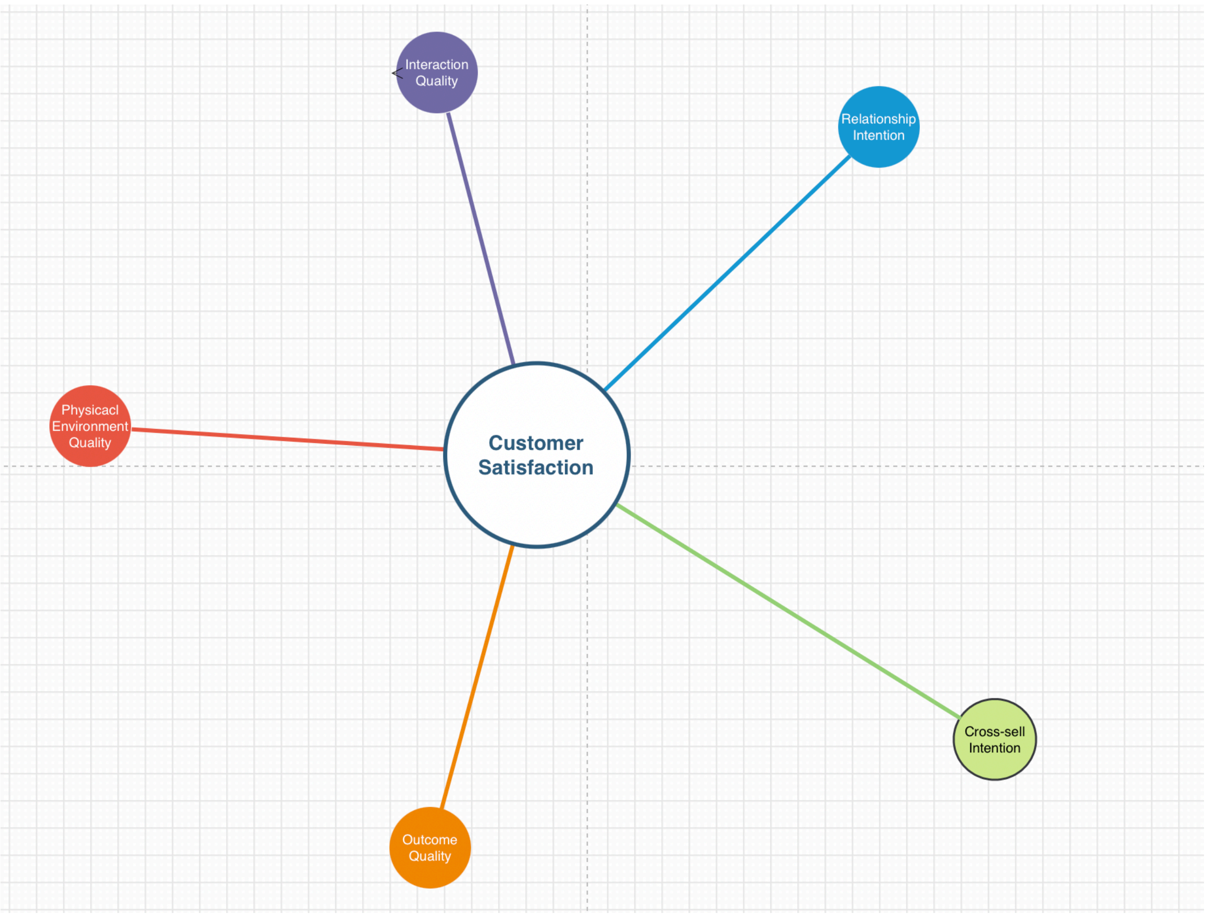


Diagram1: Connection of customer satisfaction

The data analysis process, or alternately, data analysis steps, associates assembling all the information, proceeding with it, examining the data, and helping it to discover patterns and other insights (*Kelly 2022*).

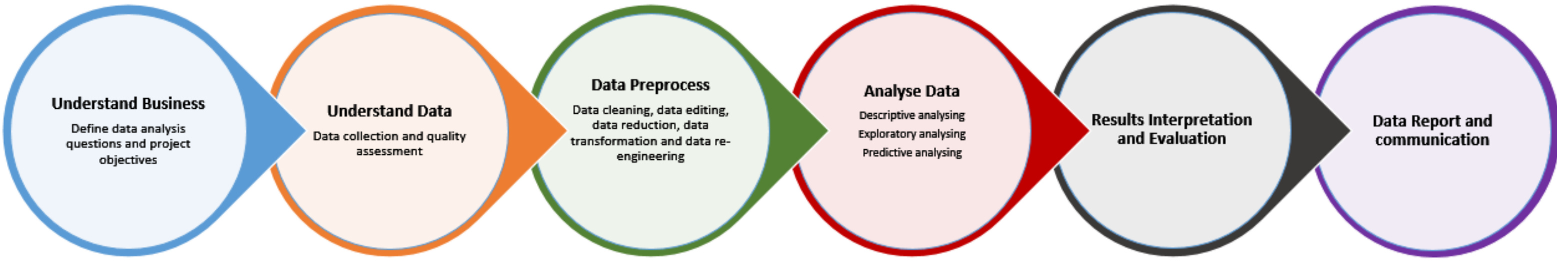

Understand Business
Define data analysis questions and project objectives
Understand Data
Data collection and quality assessment
Data Preprocess
Data cleaning, data editing, data reduction, data transformation and data re-engineering
Analyse Data
Descriptive analysing
Exploratory analysing
Predictive analysing
Results Interpretation and Evaluation
Data Report and communication

Figure 1: Process of data analysis (*Li 2022*)

·Step 1: Selecting data - Selecting suitable data for collection-analysis purposes. Development and implementation of technical measures to collect them.

·Step 2: Data extraction, processing, refining - Extraction of necessary data, Creation of new analytical variables, classifying text values into specific categories.

·Step 3: Data integration, summary - Integrate each data based on key variables and generate statistics.

·Step 4: Generate predictive models - Create and select high-performance predictive models about big data.

·Step 5: Interpretation and application of results - Examine a validation study of the production results of the prediction model.

·Step 6: Application and feedback - Apply analysis results directly. Modify and improve the prediction models.

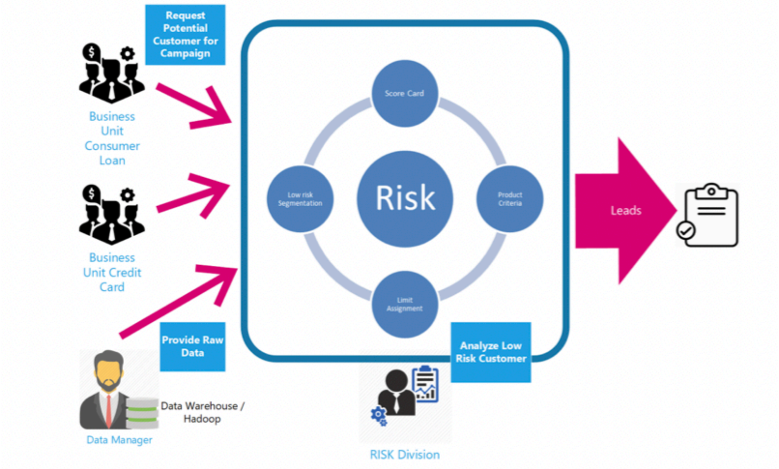


Figure 2: Current Cross-Sell Activity (*Rakhman 2019*)

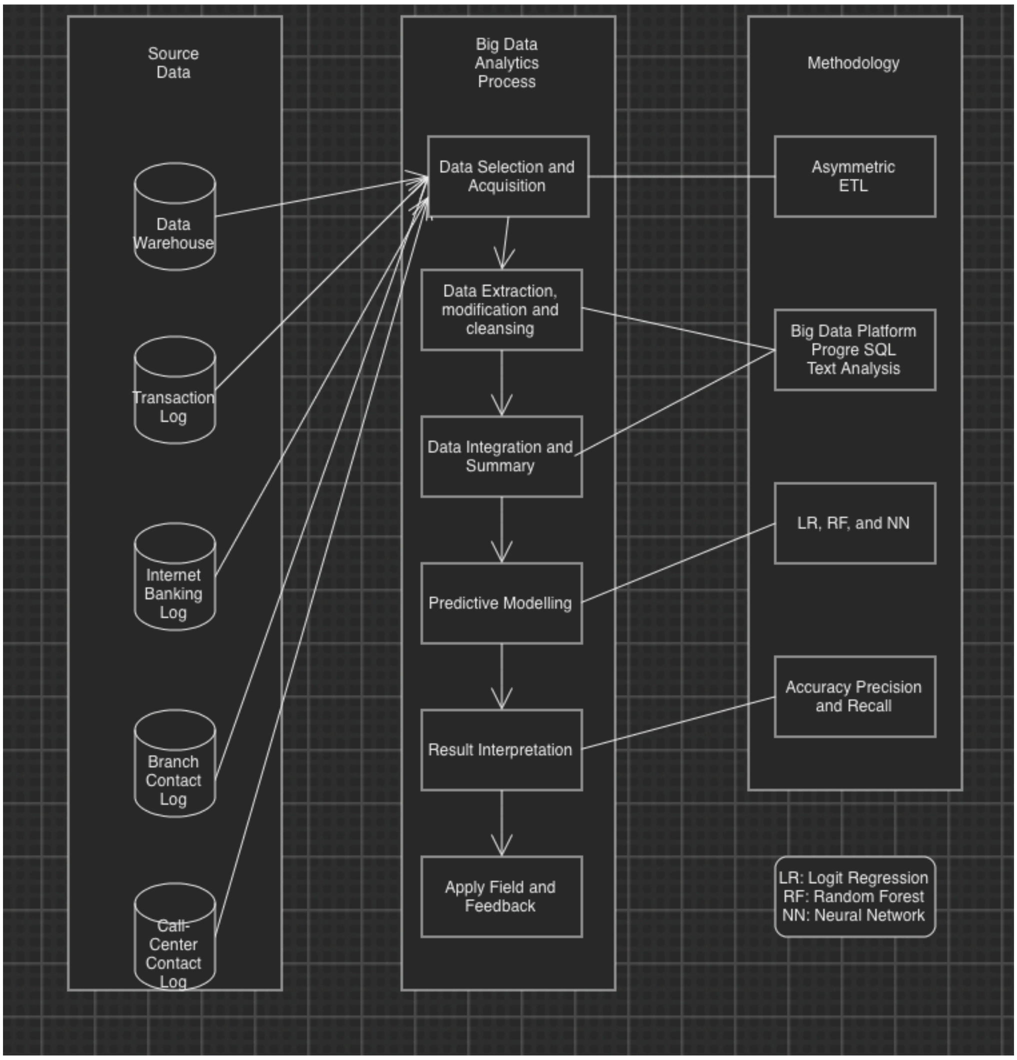


Diagram2: Diagram of Big Data Analytics for Target marketing in Banking

## Data sources

The detailed definition and contents of 5V are as follows

1. Large Capacity - It usually means a large amount of data that uses a large amount of storage or consists of a large number of records. Banks generate a variety of vast amounts of data in goods and systems such as money entry and withdrawal and transfers, deposits, consumer loans, commercial loans, mortgages, automated device transactions (ATMs), Internet banking, and mobile banking logs. To store and manage this data, banks are mainly using relational database management systems (DBMS). However, for the processing and analysis of big data, distributed/ parallel processing-based analysis platforms such as Hadoop and Map Reduce are required.

2. High Speed - It means the frequency of occurrence of data and the frequency of transmission of data. Digitalized data collection is rapidly taking place due to the development of IT technology and the increase in the use of non-face-to-face channels such as Internet banking and mobile banking.

3. Diversity - It refers to data generated in different formats from various sources of data and including multidimensional data fields. In addition to the data generated inside the bank, such as transaction records, it includes social media, terrain space data, and other public information generated outside the bank.

4. Accuracy - It is also expressed as the reliability of the data, and it means the selection of error data and the generation of relevant data. High data quality is an important requirement for applying big data analysis. Therefore, there is a need for the ability to collect verified and relevant data and block bad data.

5. Value - The importance of big data analytics as a stepping stone to the advancement of knowledge, innovation, and improved decision-making processes should be recognized. Taking into account the perspective of information benefits when establishing a strategy for big data, creating value by better utilizing big data assets in addition to business assets and human resources is the basis for securing a competitive advantage for companies.

At the beginning of the project, data to be collected were selected through a consultation agreement between data scientists and marketers. Unstructured external data such as social data were judged to consume more time and cost in data collection and processing compared to the effect that can be obtained so they were excluded from the collected data. In order to apply big data analysis inside the bank quickly, the range of target data was limited to five data sources and data was collected. All data remove the customer identification information such as an address, phone number, and email address and was used for analysis only statistical information that could not be customer identified. In the data warehouse, basic customer information and monthly account history data of customers were collected through the big data platform. After that, fund transfer logs, Internet (mobile) banking logs, face-to-face contact history, and call centre counselling records were collected and transferred to the big data platform. Also, data were collected using reliability and Validity Verification and give a detailed analysis of architectures recommended for the finance company.

## Big Data architecture

| **Architecture** | **Summary** | **Technology Focus** |
|---|---|---|
| Distributed belief beween banks | Gain how to build a believed environment for details dividing without applying to a integrated database, in banks or other financial inderstries. | Blockchain |
| Copy and correspond mainframe data in Azure | Copy and correspond mainframe data to Azure for digital alteration of conventional banking systems | Mainframe |
| Improve mainframe & midrange data | End to end improvement scheme for mainframe and midrange data sources. | Mainframe |
| Refactor IBM z/OS mainframe Coupling Facility (CF) to Azure | Gain how to influence Azure services for scale-out performance and good accessibility, analogous to IBM z/OS mainframe systems with Coupling Facilities (CFs). | Mainframe |
| Banking system cloud alteration on Azure | Gain how a crucial bank improved its financial deal system while keeping compatibility with its having payment system. | Migration |
| Patterns and applications in banking cloud alteration | Gain the design patterns and applications worked for the banking system cloud alteration on Azure | Migration |
| JMeter application reference for load examining pipeline solution | Gain about an application for a capable of being scaled cloud load examining pipeline worked for the banking system cloud alteration on Azure | Migration |
| Real-time fraud detection | Gain how to find data in real time to detect fraudulent deals or other irregular activity. | Security |

table1: Analysis of architectures for finance company(*Azure 2022*)

# Architecture

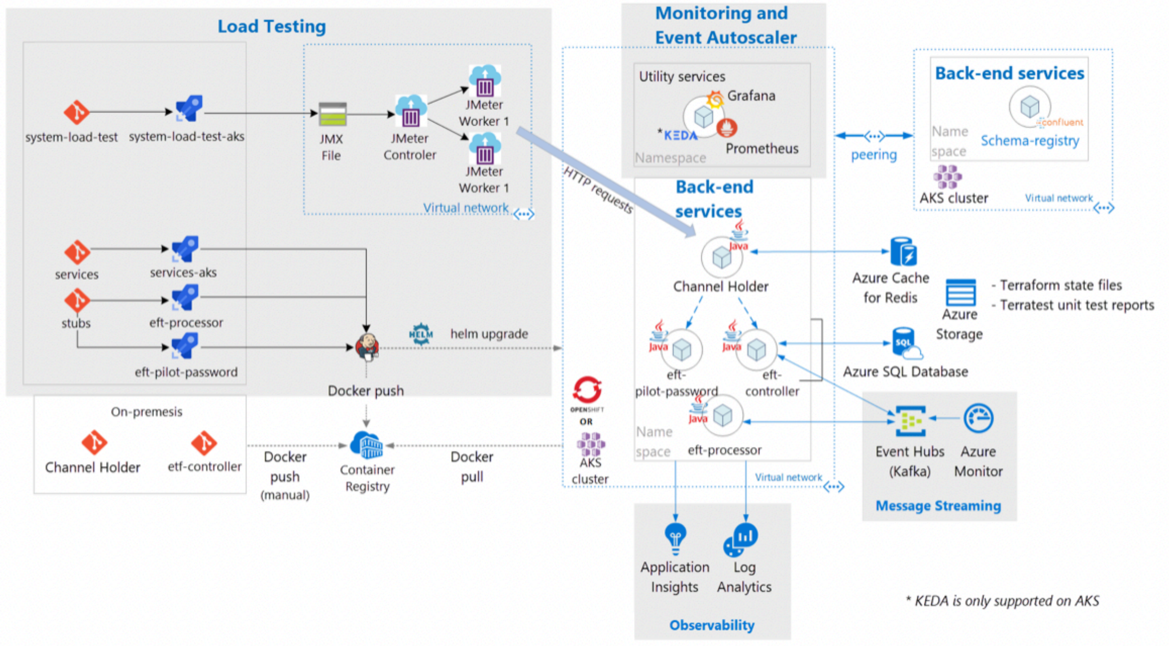


Figure 3: Architecture of Banking system cloud transformation on Azure(*Azure 2022*)

Three central blocks compose the solution: back-end services, load testing, and monitoring with Event Autoscaler (*Azure 2022*).

In the data extraction, processing, and purification stages, previous data analysis and big data analysis are typically differentiated. From the perspective of the diversity of big data, information was extracted from various data that were not previously used and analysis variables were generated. Text analysis is applied to the text of the bank record of the fund transfer log, which is unstructured data, to identify the purpose of the fund transfer. Text analysis is also called "data mining", which supports banks operating with credit ratings and anti-fraud arrangements, examining client financial data, buying deals, and card deals (*Simplilearn 2022*). Data mining also supports banks in better figuring out their clients' online customs and desires, which supports when designing a new marketing campaign (*Simplilearn 2022*). For example, it extracts information that can understand customer behaviour and statuses, such as credit card companies and payments used by customers, insurance and premiums paid, and pension receipts. Also, call centre counselling records to use text analysis to extract information about customer inquiries, complaints, and concerns. It extracts page information visited by online customers by parsing the Internet (mobile) banking log, which is semi-structured data and extracts customer behaviour patterns such as Internet (mobile) banking visit cycles, used financial services, and inquiry into financial products. The main contents of the data extraction step are shown in <Table 2>. The meaning of the explanatory variables in this case and the main variables are shown in <Table 3>. Dependent variables and explanatory variables are extracted and generated for each data source that was integrated into the form of learning in the prediction model. The analysis goal of this case is to select a target customer for sales of financial products, and variables are summarized and integrated centring on customers. Time series data such as monthly card payment amount and premium payment amount derived through text analysis were simplified and then summarized as rising, maintaining, and falling over by time flow. In this study, the prediction model has used a dependent variable to calculate the probability of product subscription. Finally, a classification algorithm is typically a role that weights the input attributes so that the output isolates one class into positive values and the other class into negative values (*Netoff 2019*). The learning model learns using statistical methods such as logistic regression (LR) and machine learning methods such as random forest (RF) and artificial neural network (NN). The calculated probability of newly collecting data to understand the marketing application effect of big data analysis and entering it into a target marketing model based on logistic regression analysis is calculated as the target customer index. The purpose of this case study is to select customers who subscribe to financial products, and the logistic regression analysis model is selected as the model with the highest Recall value, which is the ratio of matching customers with actual product subscriptions among the three prediction models. Therefore, the final target marketing model based on big data analysis will have many explanatory variables, but only a few explanatory variables generated from big data will appear as shown in <table 4>. As a result of big data analysis, some variables were able to find out customer patterns that were not recognized until. For example, the receipt of severance pay and the generation of pension income showed a positive (+) effect on the subscription of received products. This indicates that retired customers have a high willingness to join relatively safe financial instruments. In addition, customers with high use of windows, automated devices, or high employee contact in financial institutions and transaction methods had a positive (+) effect on product subscription compared to non-face-to-face financial transaction-oriented customers such as Internet banking and smart banking. This can be seen as a characteristic of the research target company that operated mainly in the traditional financial industry in the transition period when the current paradigm of financial institutions is changing to non-face-to-face electronic finance. In addition, if the paid-in insurance premium decreases, it has a positive (+) effect on the subscription of the received product, but if the paid-in insurance premium is maintained or increased, it has a negative (-) effect.

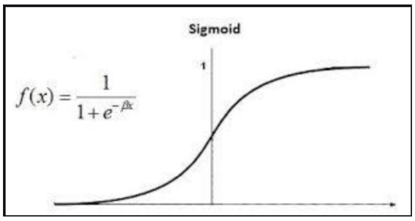


Figure 4: Logistics Regression (Rout 2020)

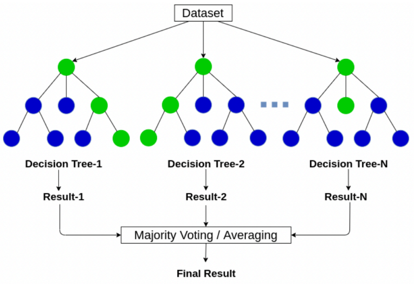


Figure 5:Random Forest(Goyal 2021)

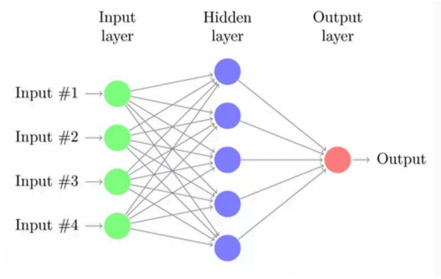


Figure 6: Neural Network(*Dilmegani 2021*)

| Data source | Methods | Extracted information | |
|---|---|---|---|
| | | Type of customer behavior | Pattern of customer behavior |
| Transaction log | Text anlaysis | Purpose of Fund Tranfer | Cycles, periods, frequency |
| Internet (Mobile) banking log | Web log anlaysis | Type of visit page | |
| Call-center contact log | Text anlaysis | Purpose of contact | |

Table 2: Extracted Data Source

| Categories | Meaning | Variables |
|---|---|---|
| Affordable to make account | Economic allowance to purchase additional items from the bank | Closeness of financial instruments expiration date |
| | | Increase amount of insurance fee |
| | | Decrease amount of insurance fee |
| | | Increase of credit card amount |
| | | Decrease of credit card amount |
| Degree of engagement for bank | Closeness between customer and financial institution | Visit or contact period |
| | | Count of contact |
| Transaction pattern | Specific patterns or trends in customer behavior during banking transactions | Count of transaction by purpose |
| | | Day count after product expiration |
| Degree of engagement for product | Closeness betwenen customer and financial instruments | Increase of contract period of financial instruments for saving money |
| | | Decrease of contract period of financial instruments for saving money |
| | | Count of fund transfer transactions |
| | | Count of financial instruments page view |
| Preference for safety asset | Safety product priority purchase | Make account when get surplus funds |
| Contact activity | Contact Behavior by channel (ATM, internet or mobile bank, off-line branch) | Count of contact channels |
| | | Type of task by channels |
| Voice of Customer | Inquiries or complaints received by the call center | Count of call center contact |
| | | Contact type |
| | | Type of customer complaint |

Table 3: Explanation of Input Variables

| Categories | Variables |
|---|---|
| Affordable to make account | · Increase amount of insurance fee |
| | · Decrease amount of insurance fee |
| | · Increase amount of income |
| Degree of engagement for bank | · Family customer |
| | · Elapsed period after contract |
| | · Elapsed period after last contact |
| Transaction pattern | · Retirement payment is received |
| | · Pension receipt |
| | · Public pension payment |
| Degree of engagement for product | · Increase of Savings product Contract Period |
| | · Decrease of Savings product Contract Period |
| Preference for safety asset | · N/A |
| Contact activity | · Number of branch visits |
| | · Count of ATM usage |
| | · Count of internet banking usage |
| | · Count of smart banking usage |
| Voice of Customer | · Ask product registration |

Table 4: Input Variables from Big Data

A confirmatory factor analysis was conducted to confirm the concentrated validity and discrimination validity of the concepts. The results of the confirmatory factor analysis for the measurement model are shown in <Table 5>, and as a result is the fitted index is x2=330.618 (d.f.=197, p= 0.000), GFI= 0. 877, AGFI= 0.851, IFI= 0.966, CFI= 0.966, RMSEA= 0.058. As a result of analyzing the concentrated validity, all standard factor load values were significant, all concept reliability (CR) was higher than 0.7, and all average variance extraction values (AVE) were higher than 0.5, so concentrated validity was secured and can be judged.

| Variable | | B | T-value | P | C.R |
|---|---|---|---|---|---|
| **Interaction Quality** | Interaction 1 | 0.817 | | | 0.856 |
| | Interaction 2 | 0.874 | 14.551 | *** | |
| | Interaction 3 | 0.901 | 15.315 | *** | |
| | Interaction 4 | 0.860 | 14.499 | *** | |
| | Interaction 5 | 0.866 | 14.669 | *** | |
| **Outcome Quality** | Outcome 1 | 0.813 | | | 0.802 |
| | Outcome 2 | 0.936 | 11.053 | *** | |
| | Outcome 3 | 0.877 | 6.878 | *** | |
| **Physical Environment Quality** | Physical 1 | 0.772 | | *** | 0.847 |
| | Physical 2 | 0.851 | 13.602 | *** | |
| | Physical 3 | 0.889 | 9.753 | *** | |
| | Physical 4 | 0.892 | 0.9133 | *** | |
| **Customer Satisfaction** | Satisfaction 1 | 0.821 | | | 0.879 |
| | Satisfaction 2 | 0.889 | 15.708 | *** | |
| | Satisfaction 3 | 0.920 | 16.503 | *** | |
| | Satisfaction 4 | 0.855 | 14.766 | *** | |
| | Satisfaction 5 | 0.554 | 8.278 | *** | |
| **Relationship Intention** | Relationship 1 | 0.710 | | | 0.812 |
| | Relationship 2 | 0.864 | 10.851 | *** | |
| | Relationship 3 | 0.787 | 10.152 | *** | |
| **Cross-Buying Intention** | Cross-Buying 1 | 0.912 | | | 0.858 |
| | Cross-Buying 2 | 0.928 | 17.6955 | *** | |
| | Cross-Buying 3 | 0.852 | 17.299 | *** | |
| x2=330.618 (d.f.=197, p= 0.000), GFI= 0. 877,AGFI= 0.851, IFI= 0.966, CFI= 0.966, RMSEA= 0.058 | | | | | |

Table 5: Analysis of Measurement Model

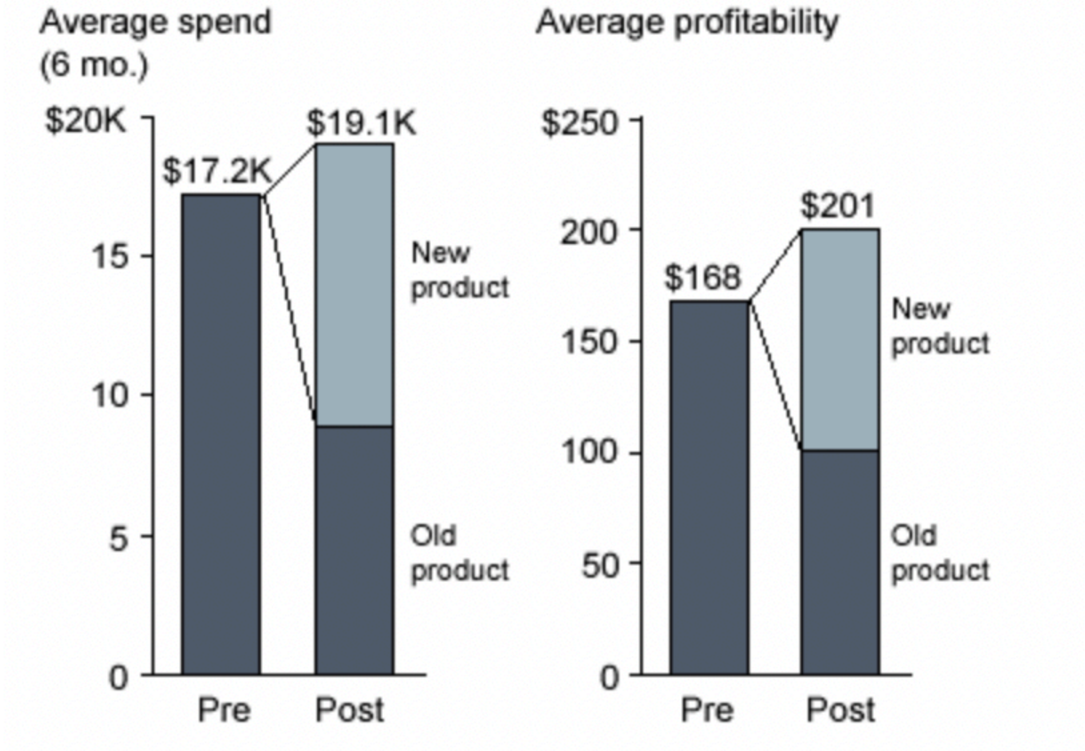


Figure 7: Cross-sell and customer relationship increasing graph(Marcil 2020)

## Benefits and challenges

The purpose of this study is to discuss the application and effectiveness of big data analysis to target marketing of banks at a more detailed level through the case of target marketing based on big data analysis in banks, present issues and implications, and present a methodology for effective application of big data analysis in the banking industry. Through this, it aims to contribute to the theory and practice of the application and implementation of big data analysis from the marketing perspective of financial institutions. This study will serve as a good example of the performance and application of big data analysis by showing the practical contents of the use of big data. First, previous studies presented partial effects using some of the various data of financial institutions, but this case study comprehensively analyzed various big data such as fund transfer logs, Internet/mobile banking logs, call centre counselling history, and customer contact history. Second, it can be confirmed that the results of big data analysis are applied to actual marketing activities, resulting in higher performance than previous marketing. The practical implications of this study are as follows. In the early stages of big data analysis, it is desirable to apply data that was not used for analysis previously, check the effect, and gradually expand the target and scope of big data. In other words, to apply big data analysis, it is effective to use vast amounts of internal accumulated data rather than external data and to gradually expand the range of data to structured and unstructured data outside the company. In addition, analysis of unstructured text and semi-structured log data is essential to extract as much information as possible from the collected big data. This information can be used as an additional explanatory variable to understand the behaviour and state of the customer that has not been previously identified, thereby being able to create a more sophisticated and robust prediction model. As for the big data analysis method for marketing utilization, 1) data selection and collection, 2) extraction and processing, 3) summary and integration, 4) prediction model generation, 5) interpretation, and 6) application and feedback are effective. The selection of data and application of big data analysis and feedback procedures included in this case study are essential for efficient application and performance measurement of big data analysis.

## Conclusion

The survey targets were mainly focused on corporate customers. Since individual and corporate customers can evaluate the service quality of banks differently, if a separate service quality study is conducted on individual customers to compensate for these problems, it will be able to present important implications to banks that manage service quality. It is difficult to say that it accurately explains the overall quality of service. In future studies, research on the appropriate scale development and refinement process for measuring factors that are more detailed and reflect customer satisfaction should be continuously conducted. Future research is needed to present recommended products that have the needs of customers targeted as a result of this study. It should be a study on customer segmentation based on big data analysis and derivation of recommended products by customer group. Also, additional research will show to construct and analyze big data by converging with other financial institutions, telecommunications, transportation, distribution, and public data using de-identification data. Through this, it is believed that more expanded customer insights can be obtained and efficient marketing and product sales opportunities can be sought. In addition, future research on the problem and importance of personal information protection, which is a prerequisite for the use of big data, will also be very important.

## References

Li, G (2021) *Do A Data Science Project in 10 Days*, Bookdown, viewed 31 May 2022, <https://bookdown.org/gmli64/do_a_data_science_project_in_10_days/>.

Developing big data solutions on Microsoft Azure HDInsight (2016) *planning a big data solution*, Microsoft, viewed 31 May 2022, <https://docs.microsoft.com/en-us/previous-versions/msp-n-p/dn749858(v=pandp.10)?redirectedfrom=MSDN>.

Rakhman, R, Widiastuti R, Legowo N, Kaburuan E (2019) *Big Data Analytics Implementation In Banking Industry – Case Study Cross-Selling Activity In Indonesia's Commercial Bank*, IJSTR, viewed 31 May 2022, <http://www.ijstr.org/final-print/sep2019/Big-Data-Analytics-Implementation-In-Banking-Industry-Case-Study-Cross-Selling-Activity-In-Indonesias-Commercial-Bank.pdf>.

Li, Y, Chiu, Y, Lin, T, Huang, Y (2019) *Market share and performance in Taiwanese banks: min/max SBM DEA*, ResearchGate, viewed 31 May 2022, <https://www.researchgate.net/publication/330969474_Market_share_and_performance_in_Taiwanese_banks_minmax_SBM_DEA>.

Kelley, K (2022) *What is Data Analysis: Methods, Process and Types Explained*, Simplilearn, viewed 31 May 2022, <https://www.simplilearn.com/data-analysis-methods-process-types-article#what_is_the_data_analysis_process>.

Simplilearn (2022) *What Is Data Mining: Definition, Benefits, Applications, Top Techniques*, and More, viewed 31 May 2022, <https://www.simplilearn.com/what-is-data-mining-article>.

Netoff, T (2019) *The Ability to predict Seizure Onset, ScienceDirec*t, viewed 1 June 2022, <https://www.sciencedirect.com/topics/engineering/classification-algorithm>.

Dilmegani, C (2021) *Dark side of neural networks explained, AIMultiple*, viewed 1 June 2022, <https://research.aimultiple.com/how-neural-networks-work/>.

Goyal, C (2021) *Bagging- 25 Questions to Test Your Skills on Random Forest Algorithm*, AnalyticsVidhya, viewed 1 June 2022, <https://www.analyticsvidhya.com/blog/2021/05/bagging-25-questions-to-test-your-skills-on-random-forest-algorithm/>.

Rout, A (2020) *Advantages and Disadvantages of Logistic Regression*, GeeksforGeeks, viewed 1 June 2022, <https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/>.

Marcil, M (2020) *Why Upselling and Cross-Selling are Essential to Businesses?*, Grabb, viewed 1 June 2022, <https://grabb.ai/2020/06/26/why-upselling-and-cross-selling-are-essential-to-businesses/>.

BBC News (2018) *Australia's Commonwealth Bank lost data of 20m accounts*, BBC, viewed 7 June 2022, <https://www.bbc.com/news/business-43985233>.

7wData (2018) *Big Data in Banking: Many Challenges, More Opportunities*, 7wData, viewed 7 June 2022, <https://7wdata.be/apache-hadoop/big-data-in-banking-many-challenges-more-opportunities/>.

Patel, H (2018) *What is Big Data Analytics and Why it is so Important?*, Medium, viewed 7 June 2022, <https://medium.com/@patelharshali136/what-is-big-data-analytics-and-why-it-is-so-important-1de86fa37540>.

Hartung, R (2017) *Data analytics drives retail banking*, theasianbanker, viewed 7 June 2022, <https://www.theasianbanker.com/updates-and-articles/data-analytics-drives-retail-banking>.

Developing big data solutions on Microsoft Azure HDInsight (2022) *solutions for the finance industry*, Microsoft, viewed 09 June 2022, <https://docs.microsoft.com/en-us/azure/architecture/industries/finance>.

Developing big data solutions on Microsoft Azure HDInsight (2022) *Banking system cloud transformation on Azure*, viewed 09 June 2022, <https://docs.microsoft.com/en-us/azure/architecture/industries/finance>.