Chapter 1
Introduction to Data Mining

ABSTRACT

Data mining is a powerful and increasingly popular tool that uses machine learning to uncover patterns in data and help businesses stay competitive. Data scientists are trained to understand business objectives and select the correct techniques for data exploration and pre-processing. After formulating the business question, data mining methods are chosen and evaluated to determine their ability to fit the data set and answer the query. Results are then reported back to the business owner. Data mining is an essential part of modern business, allowing the organization to keep up with the competition and remain successful. With its growing popularity, the need for data scientists is rapidly increasing.

INTRODUCTION

Nowadays, data analysis using data mining is becoming more and more important. The digital ecosystem supports the storage of massive amounts of data either Cloud Technology or Database with large amount of memory. Moreover, these data are generated with users around the world in real time with high the speed communication through Social Networks resulting in a massive growth of data, known as Big Data. Businesspeople around the world can take advantage of it. For instance, to analyze customer purchasing patterns or predict the number of raw materials for production, it requires Data Mining techniques to analyze the patterns and predict the outcomes necessary for entrepreneurs' decision making.

Data analysis for business decision making can be done with various techniques such as Descriptive Statistics and Inferential Statistics. Therefore, the Data Mining technique is not a replacement for traditional data analysis, but it should be considered the development of advanced analysis techniques along with the work of machine learning technology. It supports the processing of large amount of data, Big Data. Consequently, Data Mining has yielded accurate results consistent with business questions.

DOI: 10.4018/978-1-6684-4730-7.ch001

IMPLEMENTATION OF DATA MINING AND CHALLENGES

With the need for problem analysis, it has created a position to be responsible for data analysis, namely Data Scientist. However, the Data Scientists are not solely from experts in digital technology, but they can be any person who owns the data. The medical technicians, who have information on the patients' health and understand the context of medical industry, question asking and the nature of the data, can become a Data Scientist.

Many industries have employed data scientists who are not familiar with the industry or lack experiences in the data they are to analyze. Therefore, the data scientists have to consult the questioners, possibly from the management department, and the experts in order to obtain results consistent with the facts. Therefore, in the scientific work, the data can be applied in various industries as presented in the following examples.

- In the market, to determine the selling price, the consumers can see a model which consists of many sub-models. The prices vary depending on the car's components. However, the consumers can still find similar pricings in all sub-models. For instance, Sub-model 1 costs 1,000,000 baht. The 2nd Sub-model is priced at 1,600,000 baht and the 3rd Sub-model is priced 1,7000,000 baht. It is obvious that the difference between the 2nd and 3rd Sub-models is only 100,000 baht. Such a price setting is caused by dividing customers into 3 groups to suit the number of Sub-models which is three. Data mining techniques are then used to analyze the mid-prices of each customer segment of each Sub-model. Unless the data scientists analyze the data using the techniques, general businesspeople may divide the pricing into 1,000,000 baht, 1,300,000 and 1,700,000 baht for Sub-data respectively. The clustering techniques are, therefore, used to determine the price of car sales to be able to set a suitable price for each target group. And the dealers also get the most profit.
- To forecast the condominium price in Bangkok (Sunkpho and Ramjan, 2020) using data analysis, it was found that variables affecting the prices of the condominium in Bangkok are the distance from the condo to Skytrain and MRT stations, the number of rooms, the number of floors and the age of the condominiums. Data scientists use deep learning techniques to analyze such variables and the prices, and found that the smaller distance from the condo to the sky train and subway stations, the higher the prices are. The more numbers of rooms are, the lower the prices will become. The more floors the condos have, the higher the prices are. And if the age of the condominium is less, its price is high. Therefore, the condominium real estate industries can consider such variables to determine the appropriate prices of the condominium projects.
- To categorize borrowers with the ability to pay debts, Banks have to classify their borrowers. To promote bank services, the borrowers are convinced to extend their loans with attractive offers such as a lower interest rate (Refinance). Data scientists can use data mining techniques to classify the customers by analyzing various variables. Since the banks have information about borrowers such as age, income, loan duration, default rate and the amount of additional loans that have already been approved, they can classify borrowers with the ability to repay their debts, so that they can offer marketing promotions. Banks can increase their interest income further.

From the sample cases, Data scientists do not only need to develop their knowledge on data analysis with data mining techniques, but have to study on statistics, database technology, data visualization and