Answering Business Questions with Data-Driven Analytic Thinking — A Case Study of Product Sales Prediction

Jiangping Wang

***Abstract*—Data is everywhere. Data is valuable business asset. It is the basis for predictive analytics. With the immense data opportunities, data-driven analytic thinking promotes viewing business problems from a data perspective and understand principles of extracting useful knowledge from data. This paper examines the realms of data analytics thinking and how it can be utilized in answering business questions. A case study of product sales prediction is presented to demonstrate different aspects of the process.**

***Index Terms*—Data analytic thinking, data mining, business questions.**

I. Introduction

The broad availability of data has led to increasing interest in methods for extracting useful information and knowledge from data, which land to the area of data-driven analytic thinking. With the immense data opportunities, the approach of data-driven analytic thinking promotes viewing business problems from a data perspective and understand principles of extracting useful knowledge from data. A data perspective will provide business people with structure and principles, and this will define a framework to analyze such problems effectively.

Data-drive analytic thinking is an approach to data processing and analytics that goes beyond traditional operational and data warehousing databases to address business problems. There are different data analytics techniques and data mining algorithms that work well on different types of data answering different business questions. Data analytic thinking brings with many opportunities to extend and enhance the value of data for business in many industries. The approaches in data analytic thinking embrace the challenges from complex data in complex context [1].

II. The Process of Data Analytic Thinking

The fundamentals of data-driven analytic thinking cover the concepts about how data analytic thinking aligns with the business and its competitive advantage, the process of data analytical thinking that helps the business to apply appropriate methods on appropriate data, and the techniques for extracting knowledge from data. It is an interdisciplinary field of data science of understanding and extracting valuable insights from data [2], as shown in Fig. 1.

Data-driven analytic thinking encompasses principles, processes, and techniques for understanding business through data analysis and data analytics. It covers a set of fundamental principles that guide the extraction of knowledge from data. Its techniques implement the principles of data science in the areas such as statistics, database querying, data warehousing, online analytical processing (OLAP), regression analysis, machine learning, and data mining.

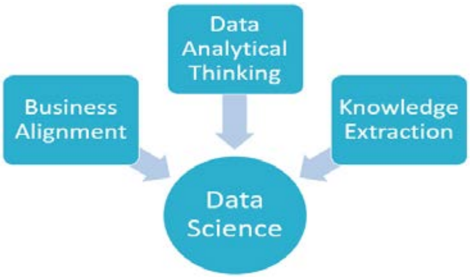


Fig. 1. Data-driven analytic thinking fundamentals.

Armed with data analytics principles, data analysts address business problems and pick the right problems that have the most value to the organization. In addition to collecting and reporting on data, they look at the data from many angles, determine the means, and recommend approaches to apply the data to the problem. Fig. 2 shows the general phases of data analytic thinking process [3], [4]. The figure illustrates its iterative nature that characterizes the process where data exploration is repeated until right understanding and right model fit right business problems [5].

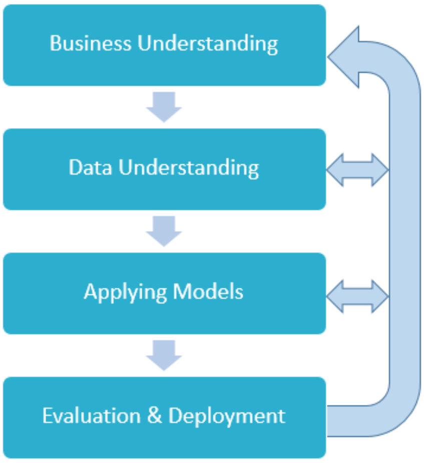


Fig. 2. Data-driven analytic thinking process.

Manuscript received March 5, 2017; revised May 1, 2017.

Jiangping Wang is with the Math & Computer Science Department, George Herbert Walker School of Business & Technology, Webster University, St. Louis, MO 63119, USA (e-mail: wang@webster.edu).

doi: 10.18178/ijimt.2017.8.3.733

III. Business Understanding

Data-driven analytic thinking starts with business understanding. If we do not have good understanding on the business and business questions we are trying to address, we will have no idea what we will need from the data, how to understand our data and, of course, how we can perform the other steps in the process of analytics.

In this information age, rich data does not automatically guarantee rich knowledge. That is why people need to mine and extract knowledge from the data to support business decision making. As a result, business understanding is vital. There are many different ways in data mining, statistics, machine learning, business intelligence, or data management, to help data understanding and knowledge extraction. The areas of data mining modeling techniques include prediction & classification, discovering relationships among entities, forecasting time series, and mining social media data. They have different features and requirements such as the goal, data preparation, analytics methods, and result applications. Matching these tools and techniques to the right business understanding is the key to success in data analytics process.

Data mining process, and in general data-driven analytic thinking, is not limited in descriptive reporting. Data description is just the first step in data understanding. Managers and executives need to answer many different types of business questions. Some are listed below:

- What is present customer base?
- What are top product groups?
- What are the sales measured per customer, per product?
- How much did new product generate by customer demographic?
- Give me sales statistics by products, by product categories.

Most of these questions can easily be answered by querying existing database with historical transactional data.

However, some other types of questions may need be addressed by more complex data analytics and data mining approaches. For example, when we try to characterize certain group of customers or describe common characteristics of top products, we will need to identify patterns based on existing data and, in turn, to help decision making for the future in order to either target most profitable customers or introduce most profitable products. In addition, if we need to examine a new customer or a new product that does not exist in our database to predict their behavior or profitability, we will have to use available data to perform predictive analysis on unseen data. In this process, it is vital to understand the business question before implementing appropriate process and techniques to improve the chances of success of data-driven decision-making.

IV. Data Understanding

Before mining any data, data analysts should fully understand their available data and know what the data represents. Data exploration in graphical visualization serves the quick way for identifying data quality and any needs in data preprocessing. Data visualization is the best way to convey the information of the data in graphical presentations. Many simple graphics can reveal rich messages, especially to the business users.

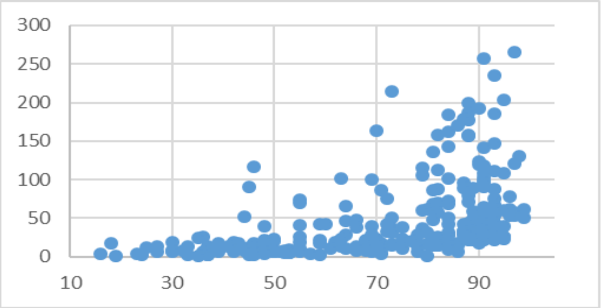


Fig. 3. Sales vs. temperature.

Here is an example of a sales prediction case where data was captured over the transactions of past year. The data includes daily sales together with some environment related data, such as temperature and precipitation. First impression from data exploration confirms that higher temperature drives higher sales, as shown in Fig. 3. The scattered plot presents only sales vs. temperature, ignoring all other factors, and each data point is treated as a standalone measure.

In addition, the U-shaped sales distribution over the seven days of a week proves higher weekend sales than weekday sales, as shown in Fig. 4. Because of that, one proposal from the user was to combine days, such as with the approach of merging Monday to Thursday into weekday category and Friday to Sunday into weekend category. However, further data analysis does not support that proposal. Since each day in a week shows sales in notable differences from each other, it might be better to leave seven days as their own category to keep the significant information in the days, instead of combining weekday and weekend as some other applications do. This is especially true if the business question will be pursued as a time series problem where valuable seasonality pattern is critical for effective prediction.

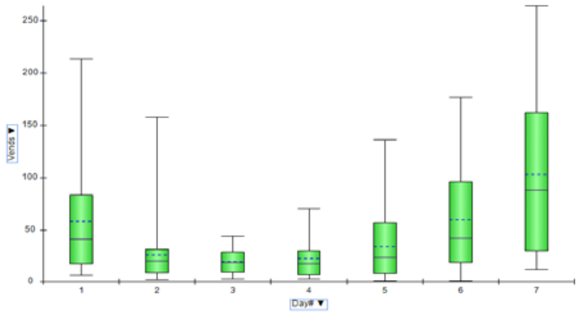


Fig. 4. Sales vs. days in week.

In terms of sales vs. precipitation, it does not seem to be the case where precipitation has too much influence on the number of sales. The variable displays not significant enough to play major role in the prediction. It was reported that in the weekends where there was rain, the sales indicated a noticeable difference compare to those on normal weekend days. To address that, we can look into weekend days separately to see the impact of precipitation. In any way, adding precipitation as one of the predictors for non-time series algorithms may capture the rain impact for all days including both weekday and weekend.

Another different approach to explore this data can be performed through the techniques in time series. Data in time series involves values that are recorded according to a specific predefined sequence of time points. Time series takes time into consideration and measurement sequence over time is important, where time is a required variable to describe the data and further to perform time series forecasting.

The date as one of the variables in the product sales case was captured in the data. The time flow of this time series of over 200 observations displays decline trend and seasonality by week, as depicted in Fig. 5. The trend can be either linear or exponential with the latter being more representative. It is expected that the sales will increase in spring months and the yearly pattern will repeat. In that case, we may have yearly seasonality on top of weekly seasonality, though more data is needed to support that assumption.

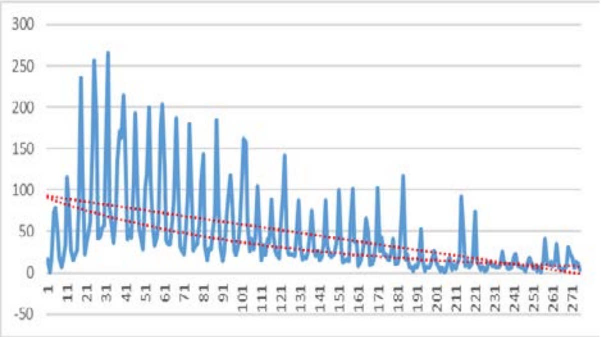


Fig. 5. Sales in time series.

V. Iterative Business Understanding

The prediction model is a generalized abstraction of the data that can be used for prediction to answer business questions and solve business problems. Data models have to build from the data and depict any knowledge from the data. By fitting a model to data, the model is developed to better interpret data characteristics and, most importantly, to decode the knowledge and serve the goal of predictive decision-making. The process of fitting a model to the data is the progression of training the model with the available data so the best set of predictors for the model can be chosen to optimize the objective.

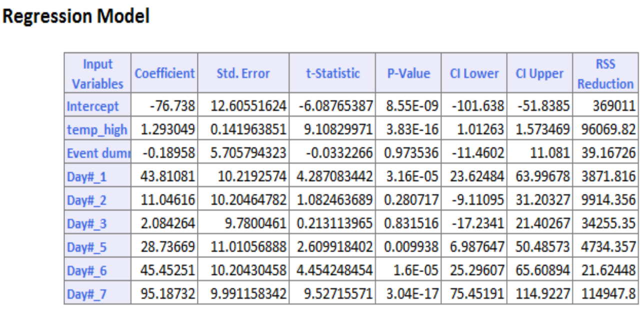
In the case study, after preliminary data exploration, it is necessary to step back to phase one in data-driven analytic thinking process, which is business understanding. Appropriate prediction methods can be identified with the help of properly defined purpose of the prediction. We need to define the business questions we would like to answer from the prediction model. This involves translating the general question or problem from previous iteration into a more specific data modeling question.

For the product sales case the data represents above, if the prediction is just about any given day with variables like precipitation, temperature, and day of week, then the following prediction algorithm can be considered: Multiple Linear Regression (MLR), k-Nearest Neighbors, or Regression Tree. The target variable is *Sales*. The predictors includes temperature, precipitation, and dummy days for all seven days of week.

For example, the MLR model finds coefficients that minimize the sum of squared deviations between the actual sales and their predicted values based on the model as defined by equation in (1).

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p \quad (1)$$

The output of the model from training data is presented in Fig. 6, where all coefficients can be applied directly in the above regression equation for prediction. At this point, we have no data for location. When more data on location is available, it can be incorporated easily in the model as a new predictor.



Regression Model

Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	-76.738	12.60551624	-6.08765387	8.55E-09	-101.638	-51.8385	369011
temp_high	1.293049	0.141963851	9.10829971	3.83E-16	1.01263	1.573469	96069.82
Event dumi	-0.18958	5.705794323	-0.0332266	0.973536	-11.4602	11.081	39.16726
Day#_1	43.81081	10.2192574	4.287083442	3.16E-05	23.62484	63.99678	3871.816
Day#_2	11.04616	10.20464782	1.082463689	0.280717	-9.11095	31.20327	9914.356
Day#_3	2.084264	9.7800461	0.213113965	0.831516	-17.2341	21.40267	34255.35
Day#_5	28.73669	11.01056888	2.609918402	0.009938	6.987647	50.48573	4734.357
Day#_6	45.45251	10.20430458	4.454248454	1.6E-05	25.29607	65.60894	21.62448
Day#_7	95.18732	9.991158342	9.52715571	3.04E-17	75.45191	114.9227	114947.8

Fig. 6. MLR model output.

It is worth to mention that the performance of various algorithms may not be the same. Therefore, evaluating predictive performance is necessary, especially against validation and test data. The supervised learning for prediction as discussed above is trained about the relationship between predictors (temperature, precipitation, day of week) and the target variable (*Sales*). After evaluating the performance from various models, the best-performed model can then be used to predict the sales in the cases where the target variable is unknown.

Time Series with trend (linear or exponential) and seasonality can be implemented if the prediction is for the near future – next few days or weeks on the time sequence immediately after the available data. The farther away we are from the available data, the less capable we can be in prediction. Therefore, lack of "randomness" might be a disadvantage. The benefit of time series forecasting is its capturing trend and seasonality. Fig. 7 presents an example of regression based time series model with exponential trend and seasonality.

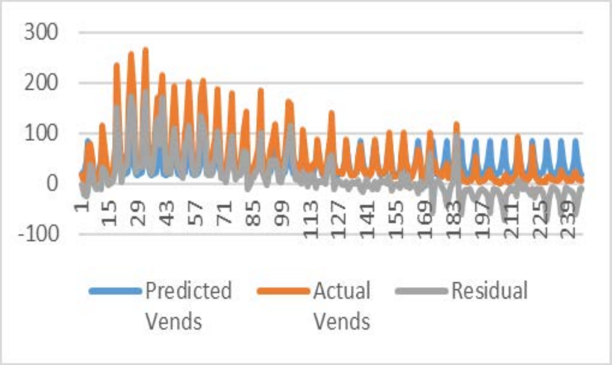


Fig. 7. Time series forecasting model.

The above time series model is built from the data and can be used to perform prediction from the relevant predictor information that is incorporated into the regression equation. With that, if the business users are clear about their business questions, they can move further along the direction they would like to pursue.

Data is the fact. However, many times the data does not automatically reveal the truth. Ultimately, data and data analytic techniques are the tools and need be used appropriately to answer business questions. If the data analysts failed to understand the data correctly, or failed to interpret the business questions, or failed to apply the right technique and right model, even the decision was made based on the data, the data will not do its job right. That is why not all data-driven decisions lead to business success. Data is a tool in the hands of users and can be interpreted as users choose. Data serves as the basis for predictive analytics and we have to see it from all different viewpoints to keep the picture crystal clear. Data was collected from business events that represent the history of business transactions. The goal of data analytics is to extract knowledge from these historical data to help our prediction about unknown future and help our decision-making about future events.

VI. Conclusion

Business problem is a business problem. This sounds so obvious. However, when we try to solve a business problem with data analytics and data mining technology, we may lose our focus, which is especially true for people in technical teams. It is crucial to understand that business problem and any associated data need be prepared to be able to fit into any technique we would like to use to attack the problem. In this process, data-driven analytic thinking can be helpful. It is important to be crystal clear on the business problems and be clear the role data analytical technique can play so the gap between the two sides can be minimized.

Data analytic principles and techniques help sift through all forms data to discover hidden insights and provide a competitive advantage in addressing business problems. Applications in all the fields of data analytics will certainly maximize information insights that business need. Data analysts and business owners can expect advancement in techniques with improved capability to tackle challenges in the age of data and age of rich information.

References

[1] H. V. Jagadis, J. Gehreke, A. Labrinidis, Y. Papakonstantinoue, J. M. Patel, R. Ramakrishnan, and C. Shahabi, “Big data and its technical challenges,” *Communications of the ACM*. vol. 57, no. 7, pp. 86-94, 2014.

[2] F. Provost and T. Fawcett, “Data science and its relationship to big data and data-driven decision making,” *Big Data Journal*, vol. 1, no. 1, pp. 51-59, 2013.

[3] D. McGilvray, “Executing data quality projects,” *Morgan Kaufmann Publishers*, Amsterdam, 2008.

[4] M. Chisholm, “7 Phases of a data life cycle”, *Information Management*, July 9, p. 4, 2015.

[5] C. Shearer, “The CRISP-DM model: The new blueprint for data mining”, *Journal of Data Warehousing*, vol. 5, no. 4, pp. 13-22, 2000.

Jiangping Wang is an associate professor of computer science at Webster University. He has received a B.A. from Chongqing University, China, a M.S. from the University of Leeds, United Kingdom and a Ph.D. from the Missouri University of Science and Technology, Rolla, Missouri, USA. Dr. Wang's areas of teaching include database design and applications, data warehousing, data mining, and data science techniques. His areas of research include database management systems, business intelligence, e-commerce data processing, and data science applications.