Transactions on Engineering and Computing Sciences - Vol. 13, No. 02
Publication Date: April 25, 2025
DOI:10.14738/tmlai.1302.18289.
Wang, J. (2025). Bridging the Gap between Business Practice and Data Science Approaches. Transactions on Engineering and Computing Sciences, 13(02). 01-09.

SERVICES FOR SCIENCE AND EDUCATION

Bridging the Gap between Business Practice and Data Science Approaches

Jiangping Wang
School of Business and Technology, Webster University, USA

ABSTRACT

In data science and analytics, the driving force is not on how to perform analytics tasks or how to use advanced technology in analytics projects. Business problems and goals should always drive the overall approaches. Projects and applications in data science and analytics should serve business goals and help business decision making. In this paper, a case study that serves various directions in answering business questions is presented.

Keywords: Data Analytics, Data Science, Big Data, Business Analytics, Decision Making.

INTRODUCTION

Data is valuable assets for business. In our time, data is everywhere. Business and organizations rely on data. Data represents business that were collection and accumulated over the entire process of business operation. As more and more data become available, business faces both challenges and opportunities of big data. Challenges are due to the characteristics of big data. [1] Data volume becomes bigger in vast amount. Data is more complex and in more heterogeneous formats generated from various sources. In addition, due to advances in technology, data are produced and collected in much faster pace, which in turn causes more data incongruency and incompleteness. Conversely, opportunities exist for business to utilize big data and enhance business decision making. [2] Data science and data analytics encompass techniques and approaches that can deal with enormous amount of complex and dynamic data from all areas of business process. [3] Analytics approaches can be implemented to represent, interrogate, and interpret all data for better understanding on what has happened in business and what the trend is to help business grow. [4] In this process, data plays empirical role. [5] Understanding business is even more important to achieve a successful data science project. There are many areas of technologies and techniques to support projects in data science and data analytics, ranging from database and data warehousing, statistics analysis, as well as methods in supervised and unsupervised machine learning. [6] Each area has unique usage in data and advantages in analysis. However, they have to be applied for correct analytical goal to serve correct business problems. Data is collected, accumulated, and manipulated in business. It represents business operations and reflects business performance. The same set of data can be examined and used to solve different business problems working towards diverse business goals. In this paper, a dataset is used to demonstrate this diverse usage for serving various business goals.

BUSINESS SCENARIO

The following business scenario is used as a case study in analysis. The data of West Roxbury includes information on single family owner-occupied homes in West Roxbury, a neighborhood

in southwest Boston, MA, in 2014. [7] The adapted dataset has 14 variables and contains over 5,000 homes. The business, such as a real estate agency, would like to, based on the predictor measurements, make predictions on total value of a property or classifications as total value high (above $400,000) or low (not above $400,000). In this way the business will be able to predict the profit, assuming higher valued houses generate more profit. The data dictionary describing each variable is available below in Table 1.

Table 1: Variable description

Variable	Description
TOTAL VALUE	Total assessed value for property, in thousands of USD
TAX	Tax bill amount based on total assessed value
LOT SQFT	Total lot size of parcel in square feet
YR BUILT	Year property was built
GROSS AREA	Gross floor area
LIVING AREA	Total living area for residential properties (ft^2)
FLOORS	Number of floors
ROOMS	Total number of rooms
BEDROOMS	Total number of bedrooms
FULL BATH	Total number of full baths
HALF BATH	Total number of half baths
KITCHEN	Total number of kitchens
FIREPLACE	Total number of fireplaces
REMODEL	When house was remodeled (Recent/Old/None)

For our analytics goals, the variable CAT.VALUE (categorical value) is added, which is derived from TOTAL.VALUE of the data indicating two categories: CAT.VALUE = 1 (above $400,000) and CAT.VALUE = 0 (not above $400,000). Dataset dimension and structure is shown in Figure 1.

```
> str(housing.df)
'data.frame':   5802 obs. of  15 variables:
 $ TOTAL.VALUE: num  344 413 330 499 332 ...
 $ TAX        : int  4330 5190 4152 6272 4170 4244 4521 4030 4195 5150 ...
 $ LOT.SQFT   : int  9965 6590 7500 13773 5000 5142 5000 10000 6835 5093 ...
 $ YR.BUILT   : int  1880 1945 1890 1957 1910 1950 1954 1950 1958 1900 ...
 $ GROSS.AREA : int  2436 3108 2294 5032 2370 2124 3220 2208 2582 4818 ...
 $ LIVING.AREA: int  1352 1976 1371 2608 1438 1060 1916 1200 1092 2992 ...
 $ FLOORS     : num  2 2 2 1 2 1 2 1 1 2 ...
 $ ROOMS      : int  6 10 8 9 7 6 7 6 5 8 ...
 $ BEDROOMS   : int  3 4 4 5 3 3 3 3 3 4 ...
 $ FULL.BATH  : int  1 2 1 1 2 1 1 1 1 2 ...
 $ HALF.BATH  : int  1 1 1 1 0 0 1 0 0 0 ...
 $ KITCHEN    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ FIREPLACE  : int  0 0 0 1 0 1 0 0 1 0 ...
 $ REMODEL    : chr  "None" "Recent" "None" "None" ...
 $ CAT.VALUE  : num  0 1 0 1 0 0 0 0 0 1 ...
```

Figure 1: Dataset dimension and structure

For the case study, business problem is to make profit in real estate market. The business would like to predict house value to estimate profit. In addition, since higher valued houses means more profit, the business would like to identify high valued houses for better buying and selling in order to make more profit.

ANSWER BUSINESS QUESTIONS

Business problems are situations where business might experience difficulties and challenges in their operations. Business would like to improve their operations by identifying and

addressing issues for better performance. These problems can be in any nature of strategy, service, people, or processes. To address these problems, data can be used for analytics project and applications for better business decision making. Understanding and navigating business problems are the starting point of implementing changes to processes, operations for efficiency and effectiveness. Identifying correct business goals for project in data analytics can help implement the applications successfully. Business problem and business goals are purely business oriented. They do not tie to any technology. Many times, it is a tendency for data science team to think them through technology perspective, which is inappropriate. Analytics goals serve business goals, instead of other way around. At the end, data can be effectively support business decision making if analytics approaches are engaged correctly.

Business problems can be addressed by accurately answering business questions. The process of answering business questions is the process of problem analysis and business understanding. It also involves understanding the differences in technological approaches and what technologies one needs to employ. Business questions and associated decision making can be addressed in many ways. Data science and analytics is one of the many that is the area our focus on.

For the given business scenario, the following sample business questions are to be answered.

- Which are the houses that have high assessed value?
- Does recent remodeling increase the probability of a house being value high?
- Can we characterize the houses that have high value?
- What value should we expect some unknown houses to have?
- Will some particular unknown houses be value high?

The first two questions are about what has happened in the past. The third question is about profiling or distinguishing observations. The last two questions are to make predictions on observations where house value or status (high value or not) are unknown.

Data science and analytics encompasses a mixture of fields and techniques in such as statistics, data query, manipulation, exploratory, analysis, as well as machine learning algorithms for predictive analytics. These questions can be answered by implementing different data science techniques. If correctly applied, various type of business questions can be answered to benefit business operations.

Which are the Houses that have High Assessed Value?

This is a type of question about what has happened. Data has been collected and stored. Data manipulation provides diverse ways in data query, aggregation, computation, summation, as well as categorization. For example, to answer the question on which are high valued houses, the following data query can be issued.

From the data, the assumption is that high valued houses are the houses that have values over $400,000. In querying data, a query condition “TOTAL.VALUE > 400” can be specified, as presented in Figure 2.

```
> # which are the houses that have high assessed value (above $400,000)?
> dim(housing.df[housing.df$TOTAL.VALUE>400,])
[1] 2212   14
> housing.df[housing.df$TOTAL.VALUE>400,c(1:8)]
     TOTAL.VALUE  TAX LOT.SQFT YR.BUILT GROSS.AREA LIVING.AREA FLOORS ROOMS
2         412.60 5190     6590     1945       3108        1976    2.0    10
4         498.60 6272    13773     1957       5032        2608    1.0     9
10        409.40 5150     5093     1900       4818        2992    2.0     8
14        575.00 7233    12288     2004       4616        2378    2.0     9
24        414.70 5216    12972     1892       3796        2054    1.5     6
41        431.50 5428     6733     1990       2880        1792    2.0     7
46        490.70 6173     5683     1995       4100        2640    2.0     6
```

Figure 2: Data query

In querying data, the search is limited by specifying search conditions. As can be seen that that data contains 2212 out of 5802 houses that have values above the specified value. This type of question can help better understand the number of existing high valued houses. For the business in housing market, they can gauge overall level of the market and develop strategies accordingly.

Does Recent Remodeling Increase the Probability of a House Being Value High?

This question cannot be answered by simply querying data. The data does not offer the answer directly. To answer the question, it is necessary to aggregate and summarize data by using approaches in statistics and probability.

One approach is to compare the probability of houses being high value with and without recent modeling. Events for above query condition can be defined as 1 ("TOTAL.VALUE > 400" is true) or 0 (otherwise). The contingency table for events can be calculated in Figure 3.

```
> # add categorical value CAT.VALUE (1 for value above 400k, 0 otherwise)
> housing.df$CAT.VALUE <- ifelse(housing.df$TOTAL.VALUE > 400, 1, 0)
> # contingency table between REMODEL vs. CAT.VALUE
> table(REMODEL = housing.df$REMODEL,
+       CAT.VALUE = housing.df$CAT.VALUE)
         CAT.VALUE
REMODEL    0    1
  None  2940 1406
  Old    322  259
  Recent 328  547
```

Figure 3: Contingency table

Statistically, conditional probabilities can be calculated for recently modeled houses. Conditional probability calculates a probability of an event happening (value high) based on the existence of another event (recently modeled). The following two probabilities can be considered and compared.

- What is the probability of a house being value high?
- What is the probability of a house being value high given that it is recently modeled?

From the contingency table, there are 62.51% of recent remodel houses are on high value. Whereas, overall, the percentage of high valued houses is 38.12%. So, the answer to the question is yes, recent remodeling does increase the probability of a house being value high. Answering this type of questions, business can better understand the impact of certain property of houses on housing value. The question can be approached by aggregating existing data.

Can we Characterize the Houses that Have High Value?

This question is type of profiling or distinguishing high valued houses. What are the characteristics of different segments of houses? How to describe high valued houses in terms

of features that are used to describe houses? Are there any commonalities among high valued houses?

From data analytics perspective, this is a question in type of unsupervised learning. Unsupervised machine learning uses methods and algorithms to analyze data. It uncovers relationship and patterns that might be hidden in the data. It can be used to explore data and better understand data segments in terms of similarities and characteristics.

To answer the question, a clustering analysis employing k-means algorithm is performed on the housing data with selected features and variables that describe each house. Features that are related to housing value can be chosen for the analysis. Four clusters are generated, and their profile can be presented, as shown in Figure 4. The profile bar plot shows characteristics of each cluster. The sizes of the four clusters are 3421, 871, 579, and 931, respectively.

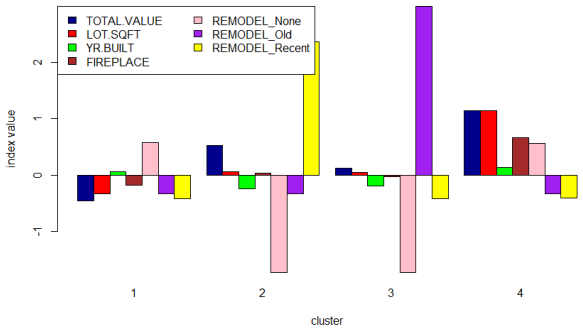

Figure 4: Clustering profile plot

Interpreting each cluster, it is noticeable that cluster 1 and 4 are distinctively different. Cluster 4 are high value houses due to big lot size and more fireplaces. Whereas cluster 1, the largest cluster, is just opposite to cluster 4 with small log size, low on fireplaces, so low house value. Values of both cluster 2 and 3 houses are increased due to remodeling. However, recent modeling (cluster 2) increases more on house value than old modeling (cluster 3).

Questions answered by such clustering analysis help business marketing strategy for reaching prospective consumers in the market and turning them into customers of their services to achieve business goals.

What Value Should we Expect Some Unknown Houses to Have?

To answer this question, a predictive model is to be built and make regression. Predictive analytics is the process of making prediction on future unknown. In this case, a house that the price is unknown would be presented to the decision maker with variables to describe the house, such as lot size, number of floors and rooms, and other features. Predictive models can be built with various algorithms on the history data to learn the relationship between predictors and the outcome variable – the house value in this case. Then the models can be used to make predictions on future unknown.

Figure 5 shows an example of multiple linear regression model and predictions on three houses. The predictors for the model include lot size, year built, cross area, and so on. The outcome variable is house total value. The first house in question (house #1) has features of 9965 lot size, built on 1880, 2436 gross area, among others. The predicted total value from the model for the first house with listed features and unknown value is $382,871.7. The predicted total value for the third house in question (house #4) is $559,178.2.

```
> # multiple linear regression
> lm_reg <- lm(TOTAL.VALUE ~ ., data = train.df[ , -c(2,15)])
> lm_reg

Call:
lm(formula = TOTAL.VALUE ~ ., data = train.df[, -c(2, 15)])

Coefficients:
  (Intercept)       LOT.SQFT       YR.BUILT     GROSS.AREA    LIVING.AREA
  -232.208413       0.008904       0.143014       0.036959       0.046462
       FLOORS          ROOMS       BEDROOMS      FULL.BATH      HALF.BATH
    43.049335      -0.357332       0.963874      16.990867      18.278777
      KITCHEN      FIREPLACE  REMODELOld  REMODELRecent
   -17.479077      19.279022       6.394886      24.216063

> holdout.df[c(1:3),-c(1,2,15)]
  LOT.SQFT YR.BUILT GROSS.AREA LIVING.AREA FLOORS ROOMS BEDROOMS FULL.BATH
1     9965     1880       2436        1352      2     6        3         1
3     7500     1890       2294        1371      2     8        4         1
4    13773     1957       5032        2608      1     9        5         1
  HALF.BATH KITCHEN FIREPLACE REMODEL
1         1       1         0    None
3         1       1         0    None
4         1       1         1    None
> predict(lm_reg, holdout.df[c(1:3),])
       1        3        4
382.8717 358.2373 559.1782
```

Figure 5: Multiple linear regression model

The model was built on the history data by learning patterns and relationship between various predictors and the outcome variable. Its prediction accuracy depends on how training data represent the relationship and how much model can learn from the data. Its performance can be tuned and evaluated by applying the model on a separate set of data. The model with satisfactory performance then can be deployed to make predictions.

The nature of this question is not about what has happened in the past. Instead, it is about future. The prediction is made based on the model that learns from what has happened on many other observations, or houses in this case study. Predictive analytics makes predictions on future and the predictive power relies on the learning and fitting models on the training data. The approach is a supervised learning process where the relationship between predictors and outcome exists in history data and need be extracted by the learning of the model. Answering this type of questions enables business better planning and predict numbers such as profit and gains.

Will Some Particular Unknown Houses Be Value High?

As discussed earlier, value high can be defined as above certain number, for example $400,000. The answer to the question can be simply yes (above the value) or no (not above the value). To approach this type of question, a predictive model can be implemented for performing classification task.

Classification is another type of predictive algorithms for modeling and making prediction on the future unknown – the unknown categorical classification. It is same as regression in terms of supervised learning nature. However, the difference lies in the type of outcome variable.

Classification answers question on outcome categorically. For example, yes vs. no. Whereas in regression, the question is about numerical outcome.

A decision tree algorithm can be applied to this problem. The classification model is shown in Figure 6, where 1 is for yes on value high and 0 for no.

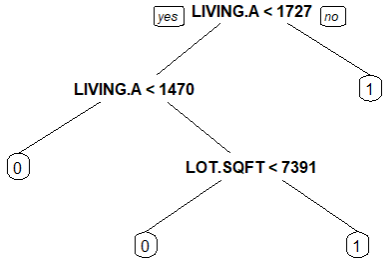

Figure 6: Classification tree model

The model represents a set of classification rules. The classification rules of the tree are transparent and easy to interpret for business owners, as listed below.

- If living area < 1470, then not above 400.
- If living area < 1720 and living area ≥ 1470 and lot size < 7391, then not above 400.
- If living area < 1720 and living area ≥ 1470 and lot size ≥ 7391, then above 400.
- If living area ≥ 1720, then value is above 400.

The classification model can be used to make classification on future houses with unknow status (above $400,000 or not). Figure 7 shows examples of classification performed by the classification tree model on the same three houses that were used in regression. The first house in question (house #1) is classified as 0 (not above $400,000). The classification for the third house in question (house #4) is 1 (above $400,000). The result of classifications on the three houses are consistent with the results of regression performed earlier.

```
> rpart_tree <- rpart(as.factor(CAT.VALUE) ~ ., data = train.df[,-c(1,2)])
> holdout.df[c(1:3),-c(1,2,15)]
  LOT.SQFT YR.BUILT GROSS.AREA LIVING.AREA FLOORS ROOMS BEDROOMS FULL.BATH
1     9965     1880       2436        1352      2     6        3         1
3     7500     1890       2294        1371      2     8        4         1
4    13773     1957       5032        2608      1     9        5         1
  HALF.BATH KITCHEN FIREPLACE REMODEL
1         1       1         0    None
3         1       1         0    None
4         1       1         1    None
> predict(rpart_tree, holdout.df[c(1:3),], type = "class")
1 3 4
0 0 1
Levels: 0 1
```

Figure 7: Classification on houses

Answering questions in nature of classification can directly be transformed into business actions. For example, since the case study assumes that higher valued houses generate more profit, a real estate agency can quickly take the action of buying or selling a high valued house for profit.

Classification, same as regression approach, is one of predictive analytics and makes predictions on future unknown. The prediction here is a categorical decision, instead of on numerical values. Classification is more valuable than regression in terms of decision-making

because in business world all the decisions made will involve some type of action taking. The question of regression in the case study may predict a house value, from which business eventually need to decide what the correct action is, or what to do, in terms of buying or selling or renting a house for profit. Proper action from timely accurate decision making helps business generate more profit from the housing market.

As a supervised learning analytics approach, classification algorithms learn from past data and generalize patterns from what has happened to form models that can be used in classification on what will happen in future. In the question, the status of house value high or not is unknown. It is possible to make prediction from the classification tree model because the model has gained insights and learned relationships from historical data.

In supervised learning, data is available in which the value of the outcome of interest (e.g., house value high or not) is known. Whereas unsupervised learning approaches are those used where there is no outcome variable to predict or classify. Hence, there is no learning from cases where such an outcome variable is known. Cluster analysis as demonstrated earlier is an unsupervised learning method where the algorithm finds clusters and segments among houses so within each cluster, there exists houses with similar characteristics and the business could plan different strategies to the group.

CONCLUSION

As presented above, there are diverse approaches in data manipulation and data analytics. All can be valuable in assist business decision making. However, it is important to understand that business problems and goals are the ultimate driving factors when choosing different methods or approaches. Before diving into any data-driven project and implementing analytics, data scientists and data analysts need to understand business and the data that represent the business. Data science and analytics is a process of interrogating and analyzing complex data and targeting to understand patterns in the data. It provides predictions in terms of regression and classification on unseen observations. In this process, historical data helps model building, identify prediction (numeric or categorical), and making decision. Predictive modeling learns from data to build models and performance can be evaluated on the data before models can apply on future houses to estimate profit. Data is valuable assets of business reflecting business operations and measures. At the same time, business face constant challenges in dynamic environment and need to make improvement based on decision making. Data-driven decision making is one of the areas that can help business in taking challenges by utilizing data and targeting improvement goals. In this process, better understand business goals is the key in order to select correct technologies and approaches to better serve business owner for decision making.

References

[1]. H. E. Brady, “The challenge of big data and data Science,” Annual Review of Political Science, 2019. Vol. 22, p. 297-323.

[2]. M. N. O. Sadiku, P. O. Adebo, and S. M. Musa, “Big data in business,” International Journals of Advanced Research in Computer Science and Software Engineering, 2018. Vol 8.1, p. 160-162.

[3]. C. Janiescha, B. Dinter, P. Mikalef, and O. Tona, “Business analytics and big data research in information systems,” Journal of Business Analytics, 2022. Vol. 5, p. 1-7.

[4]. A. C. Ikegwu, H. F. Nweke, C. V. Anikwe, U. R. Alo, and O. R. Okonkwo, “Big data analytics for data-driven industry: a review of data sources, tools, challenges, solutions, and research directions,” Cluster Computing, 2022. Vol. 25, p. 3347-3387.

[5]. A. C. L. Ziora “The role of big data solutions in the management of organizations. review of selected practical examples,” Procedia Computer Science 65, 2015. p. 1006–1012.

[6]. O. O. Amoo, F. Usman, E. S. Okafor, O. Akinrinola, and N. A. Ochuba, “Strategies for leveraging big data and analytics for business development: a comprehensive review across sectors,” Computer Science & IT Research Journal, 2024. Vol. 5(3), p. 562-575.

[7]. City of Boston Assessing Department, “Property assessment FY2014,” https://data.boston.gov/dataset/property-assessment, accessed September 2024.