Principle of data science and big data

Big data is vital for businesses in the digital age as our lives become increasingly obsessed with information. Monetizing this information has become the focus of almost every organization. Pattern recognition and forecasting techniques are giving new dimensions to business capabilities. For example, product recommendation systems benefit both sellers and customers. This will make it easier for shoppers to access the items they are looking for and increase the opportunity for sellers to earn higher profits.

But big data is only one piece of the puzzle. From the overview, you will see it is overlapping with data science, data analytics, and artificial intelligence. They are all related to statistics and mathematics branches that enable pattern recognition and predictive capabilities. There are high demands for data literacy in engineering, business, medical, and other related fields.

Why data science?

For instance, you may gather data each time customers visit your website or store. Such data might include items they add to their cart, how and how long they complete their purchases, which items customers usually purchase in bundles, and how they engage with a social media post. After ensuring the data from each source is accurate, you can identify trends in their purchasing behaviors.

Understanding your customers and what drives them to buy can help you ensure that you offer products that meet your target customers' needs. And to make sure your marketing and sales strategies are in a way that is suitable for the competitive environment. Having reliable customer data can help the business retarget, create personalized experiences, and improve the website to suit customer behaviors. With data science, we can identify trends in the dataset, make predictions, and find the probability of each possible outcome.

Why big data?

As mentioned earlier, data science roots on statistics and mathematics. More specifically, we aim to find patterns or predict future trends from existing datasets. These pictures show a basic concept of linear regression. You can see that the bigger dataset we have, the more accurate the representation or prediction it will be.

What’s the problem?

With modern computing, data calculation tasks can be easy. However, with a massive amount of input data, even a decent computer will take some time to process. And the amount of data in the digital world is increasing exponentially. You will see that, at one point, traditional computing will become inefficient.

On the other side, people need a fast and real-time result. Investors can gain or lose their money in a matter of seconds or even milliseconds. Digital advertisements have seconds to impress customers and stop them from scrolling away.

The life of many patients relies on fast diagnostics. With more data to work with but less time to spare, you can now see one of the issues. Even though you have a pile of data, not every single piece is usable. This leads to another challenge in data science to locate and extract valuable data from a substantial wasteful dataset.

Common data processing in data science

Several algorithms can be used to process the data. Even if the calculating task belongs to someone else, as a designer or manager, you still need to know how these methods work, including their limitations.

There are four critical steps in the data science study. First, data must be processed and prepared for analysis. The next step is to select the appropriate algorithm based on various requirements. Then the parameters of the algorithm must be adjusted for optimal results. Then create a model for users to analyze, compare, and choose the best approach or answer.

Data Preparation

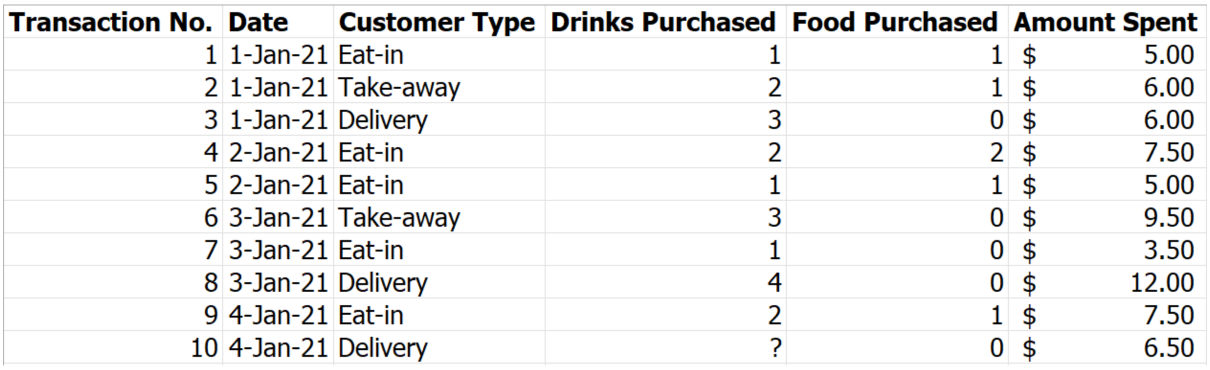
Data science is all about data. There is a simple rule called “garbage in, garbage out.” If the data quality is poor, the results will be inaccurate. This section will discuss the basic data formats that are often used in the analysis, including how to prepare data for accurate results.

Data Format

The tabular form is most commonly used to represent data. It can be found in typical databases. Each row, called 'record' or 'entry,' represents a single transaction or observation. Each column, called 'field,' shows a variable describing the data point.

Variables are also known as attributes, features, or dimensions. According to the data shown in this table, we can see some patterns across a number of transactions. If we want to investigate transaction patterns across dates, we need to represent each row as an aggregate of the daily transactions. Spreadsheet programs, like Excel, have a tool for this data rearrangement.

Table 1 – Data format



Transaction No.	Date	Customer Type	Drinks Purchased	Food Purchased	Amount Spent
1	1-Jan-21	Eat-in	1	1	$ 5.00
2	1-Jan-21	Take-away	2	1	$ 6.00
3	1-Jan-21	Delivery	3	0	$ 6.00
4	2-Jan-21	Eat-in	2	2	$ 7.50
5	2-Jan-21	Eat-in	1	1	$ 5.00
6	3-Jan-21	Take-away	3	0	$ 9.50
7	3-Jan-21	Eat-in	1	0	$ 3.50
8	3-Jan-21	Delivery	4	0	$ 12.00
9	4-Jan-21	Eat-in	2	1	$ 7.50
10	4-Jan-21	Delivery	?	0	$ 6.50

Variable Types

It is important to distinguish between different data types to ensure that they are appropriate for data analysis. Integers are used when the information must be represented as a whole number. The number of items purchased is a good example. For example, a customer cannot buy a half piece, and the number of visitors cannot be a decimal.

Continuous data represents numbers with decimal places. The amount of money spent is a good example. Other examples are weight, height, length, and speed.

A binary is the simplest variable type with only two possible options: yes or no, true or false. Any algorithm can return a binary value. For example, we can check if a customer by food items or not. The information will assist us in case we have a promotion that is valid only with a specific item type.

Data can be represented as a categorical variable when there are more than two options. The item category is an example of categorical variables in the above example. In a general customer survey, categorical variables include educational level, occupation, country, and more.

Formatted data like paragraph texts, photos, and other digital files must be encoded before they can be represented as variables for further analysis. That can be done using pattern recognition or AI.

Variable Selection

While we might be handed an original dataset that contains many variables, throwing too many variables into an algorithm might lead to slow computation or wrong predictions due to excess noise.

Hence, we need to shortlist the essential variables. Variable selection is often a trial-and-error process. Variables are taken in or out of the calculation based on the perceived quality of the results. Suppose you add one more variable into the calculation, and you feel that the result becomes less meaningful. In that case, it is a signal that the newly-added variable might be irrelevant to the problem.

Feature Engineering

Sometimes variables need to be engineered to become meaningful. From the above example, it may be difficult to predict what customers are most likely to buy food items. But if you look in the opposite direction: what type of customers do not buy food?

We find that delivery customers often do not buy food. If we regroup the variables, we might conclude that customers who purchase food items have seen the actual food they are going to buy.

Combining multiple variables is technically known as dimension reduction. It can be used to extract the most useful information and shrink it to a new smaller set of variables for analysis. However, it is difficult to draw such conclusions as the amount of data on hand is still small.

Missing Data

Sometimes data may be missing. From the previous example, the amount of food purchased in the last transaction was not recorded. It might be due to a system error, or the linked delivery platform does not provide such information. Missing data can interfere with the analysis.

We can handle missing data in one of the following ways. We can replace the missing data with the central tendency of that data set. If the missing value is a binary or categorical variable type, it can be replaced by ‘mode’ or the most extensive value.

Missing data in integer or continuous format can be replaced by the ‘median' or 'mean.' As in the example, we can replace the missing data with the median of the dataset, which is 2. We do this to minimize the effect of the missing data and avoid creating outliers.

We can estimate the missing value by referring to a similar transaction. For example, delivery customers often order 3 to 4 drinks without ordering any food item. In practice, we can compute using advanced algorithms under supervised learning which will be discussed in the next section. Although it takes longer, the estimation is usually more accurate.

If the missing value cannot be estimated, the last method we can do is to remove or exclude it from the calculation. This means deleting the entire row of that transaction. However, this method is generally avoided because it reduces the amount of data available for further analysis and may be classified as concealing or misrepresenting the transaction. In terms of data analysis, excluding calculated data could skew the trend of the remaining dataset.

Algorithm Selection

After the dataset has been processed, it can be analyzed. The choice of algorithm depends on the task we are going to perform. They can be categorized into three main categories using the AI concept.

Unsupervised Learning

When we want to find patterns in our dataset, we could use unsupervised learning algorithms. These algorithms are unsupervised because we do not know what patterns to look for. The algorithms will run recursively through all possibilities and suggest some patterns they found. From the previous example, an unsupervised learning algorithm could detect which types of customers often buy drinks and food together using association rules or cluster customers based on their purchases.

We could validate results manually by checking if customer clusters generated correspond to familiar categories. In this case, we have customers who see actual products they are going to buy and those who don't. You might not see any use of AI in this example because the dataset is too small. Typical sales records will have hundreds or thousands of records and ten or more parameters.

Supervised Learning

Supervised learning algorithms could be used when we have a hypothesis on what a pattern should look like. These algorithms are supervised because we want them to base their predictions on pre-existing patterns. In the previous example, a supervised model can predict the number of food items a customer would purchase based on their types and the number of drinks they bought.

We can directly assess the accuracy of a supervised model by adding input values from existing records and then checking how close the predictions are to the actual output. When we predict integers or continuous values, we would be solving a regression problem. This is within the capability of common spreadsheet programs and statistical apps.

When we predict binary or categorical values, we would be solving a classification problem. Most classification algorithms can also generate predictions as a continuous probability value, such as a percent chance of a particular output to occur.

Reinforcement Learning

With unsupervised and supervised learning, models are calculated based on the existing dataset and deployed without further changes. In contrast, reinforcement learning continuously updates the result as it receives additional data or feedback. The prediction will be strong if it is immune to new inputs added. Otherwise, it indicates the changes in existing trends, which should draw the manager's attention.

Other Considerations

Besides the main tasks they perform, algorithms also differ in other aspects, such as their ability to analyze different data types and the nature of results they generate. These are covered later on in the course.

Parameter Tuning

Many algorithms available in data science can create a wide variety of models. Even if the same algorithm is applied to the same set of data, the results may vary depending on the parameter adjustment. Parameters are options used to customize the algorithm's settings. It's like tuning the radio for the right frequency.

Different algorithms have different parameters to adjust. The accuracy of the model will decrease if the parameters are not adjusted properly. This example is an algorithm for classifying data types by creating a line separating two data sets.

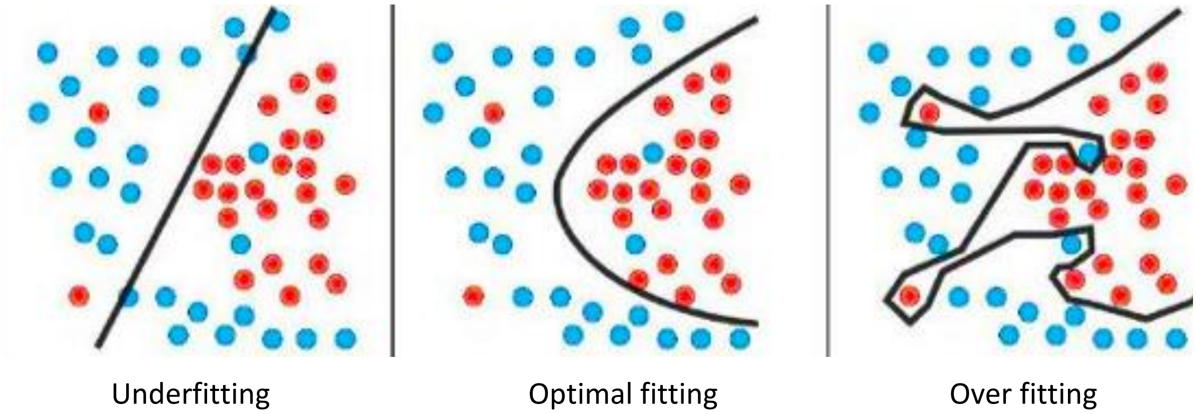


Figure 1 – Parameter tuning in classification problem
Source: Minhas, M. S., 2021

In the figure on the left, the algorithm is not sensitive enough to classify the data accurately so that many items are still mixed. This makes predictions less accurate for both current and future data.

The figure on the right shows the opposite. The algorithm is too sensitive that the dataset is divided in a non-standard pattern. That might be suitable for analyzing current data, but it does not show any trend that supports us in predicting future data.

When the parameters are properly adjusted, as shown in the middle figure, the model can accurately classify the current dataset. The separation line also shows the trend of data, so that we can use it to predict future values.

In an effort to reduce prediction errors, we might be forced to increase the sensitivity, which ultimately leads to overfitting. This happened in many data science studies. One way to maintain the overall complexity of the model is to introduce penalty parameters in a process known as 'regularization.' This new parameter penalizes the increase in model complexity by exaggerating the prediction error. This allows the algorithm to describe both the complexity and the accuracy in optimizing its original parameters. By keeping a model simple, we help to maintain its generalizability.

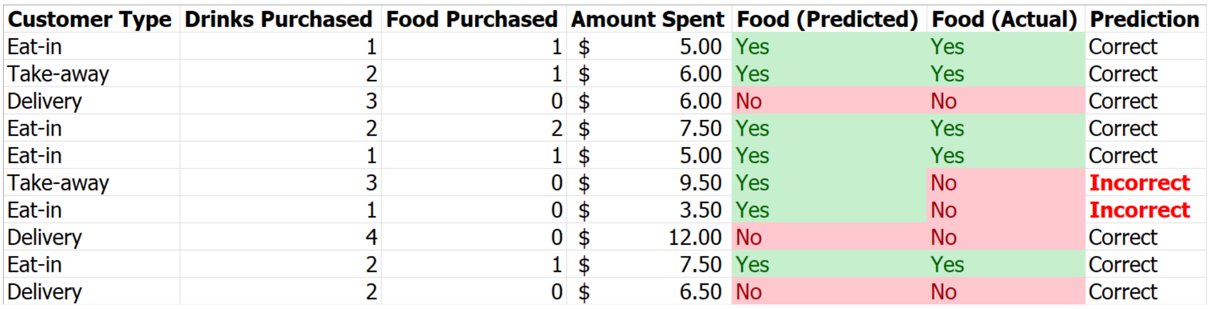
Evaluating Results

After the model is created, it must be evaluated. Evaluation metrics are used to compare how accurate a model is in making predictions.

Classification Metrics

The simplest definition of predictive accuracy is the proportion of predictions that have proven to be correct. This example shows a prediction of whether a customer will buy food items or not. Since we have historical data, we can determine whether the predictions are correct or not. Although this method is easy to understand, it doesn't really provide information about where the prediction error occurred.

Table 2 – Data prediction



Customer Type	Drinks Purchased	Food Purchased	Amount Spent	Food (Predicted)	Food (Actual)	Prediction
Eat-in	1	1	$ 5.00	Yes	Yes	Correct
Take-away	2	1	$ 6.00	Yes	Yes	Correct
Delivery	3	0	$ 6.00	No	No	Correct
Eat-in	2	2	$ 7.50	Yes	Yes	Correct
Eat-in	1	1	$ 5.00	Yes	Yes	Correct
Take-away	3	0	$ 9.50	Yes	No	**Incorrect**
Eat-in	1	0	$ 3.50	Yes	No	**Incorrect**
Delivery	4	0	$ 12.00	No	No	Correct
Eat-in	2	1	$ 7.50	Yes	Yes	Correct
Delivery	2	0	$ 6.50	No	No	Correct

The confusion matric provides further insight into where our predictive model succeeds and where it fails. This example shows a prediction of how many customers will buy food items and how many won't. Out of 100 transactions, one prediction is true positive and 89 predictions are true negative. These yield 90% overall accuracy.

Table 3 – Confusion matric

		Prediction	
		Will buy	Will not buy
Actual	Buy	**1** (True positive)	**5** (False negative)
	Don't buy	**5** (False positive)	**89** (True negative)

However, five predictions are false positives, and another five are false negatives. It can be seen that the "no-buy" forecast is much accurate than the "buy" forecast. We can also see that the prediction error is divided evenly between false positive and false negative. It is important to differentiate between prediction errors. For example, false negatives in earthquake prediction (forecasting that no earthquake will occur, but it happens) cost more than a false positive (or anticipating an earthquake but not happening).

Regression Metric

As regression predictions use continuous values, errors are generally quantified as the difference between predicted and actual values. The root mean squared error (RMSE) is a popular regression metric and is particularly useful in cases where we want to avoid large errors. This is because each individual error is squared, amplifying it on the progress. This makes RMSE extremely sensitive to outliers.

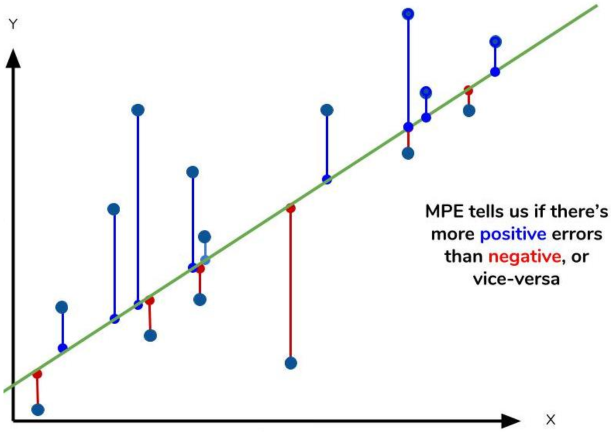


Figure 2 – Regression metrics
Source: Pascual, C., 2018

Validation

The metrics may not represent the performance of the model. Like the overfitting example mentioned earlier, the model with the lowest prediction error might not be the best answer. Therefore, we should always evaluate models using validation procedures.

Validation is the evaluation of a model's accuracy in predicting new data. We can divide our current dataset into two parts. The first will serve as a training dataset to create and refine forecast models. While the second will serve as a representation of the new data and serve as a test dataset to assess the model's predictive accuracy.

The best model is the one that provides the most accurate predictions in the test dataset. To make this validation process efficient, we should define data points to train and test datasets randomly and without bias.

However, if the original data set is small, we may not benefit from setting up test datasets because the accuracy of predictions decreases as the amount of data used to train our model decreases. Instead of using two separate datasets for training and testing, cross-validation allows us to use a single dataset for both purposes.

Cross-validation breaks the dataset into segments that will be used to test the model iteratively. This test is repeated until each segment is used as test data. Because different segments are used to generate forecasts for each iteration, the predictions will vary, and we can estimate predictiveness using the average of all iterations. If the measurement accuracy is low, we can go back and recalibrate the parameters or reprocess the data.

Summary

There are four key steps in a data science study. First, we prepare the data for the analysis. Only accurate and relevant variables should be included. As garbage cannot turn into gold, inaccurate data cannot turn into useful information. Sometimes, data need to be engineered to become meaningful. And sometimes, we have to deal with missing data by either approximating or removing them.

Next, we choose the algorithm that best fits the purpose and matches the types of data we have. We use regression for integer or continuous data and classification for binary and categorical data.

Even though we apply the same algorithm to the same dataset, the results may vary depending on parameter tuning. This process is to make sure that predictions are relevant and reasonable, minimize error, and avoid under-fitting or over-fitting.

A model with the lowest error might not always be the best answer because it can be too complicated and not show any predictive trend. Using a regularization technique can help balance error and complexity.

We can evaluate our model by comparing predicted and actual data. The overall predictive accuracy is the percentage of correct predictions or the average prediction errors.

Reference

Ng, A., & Soo, K. (2017). *Numsense! Data Science for the Layman: No Math Added.* Annalyn Ng and Kenneth Soo.