Errors, biases and algorithms: how to interpret automated results

Does data lie? Misconstrued results may be deliberate distortion so it is important to consider carefully the underlying process. ***Alex Russell*** *of Bates Group LLC sets out some examples to help fraud examiners avoid the pitfalls of working with data.*

Sense check

Machine learning and artificial intelligence are pervading all aspects of modern life, from the helpful voice assistant in your phone to the prediction algorithms that help you choose a movie or a new restaurant. The success of platforms that rely on machine learning has led to a sense of complacency about the inner workings of these automated tools. The default assumption for most people is that the programme is performing correctly, and making 'fair' decisions, free from the types of errors to which human judgement is prone. Given the built-in trust placed in these systems, it is important to understand their limitations, which are especially obvious when compared to human judgement.

Learning style

All machine learning problems can be broadly classified by the method that the machine uses to actually learn. One such is Supervised Learning, a model that includes outcome labels created by a human 'teacher'. The algorithm is then given a set of input features and trained to approximate a function that makes predictions about the output. To give a very simple example, imagine a lender that only has two pieces of information captured for every loan it has ever extended: credit score, and whether the loan was repaid in full. In supervised learning, the algorithm would take these two pieces of information (with the input feature being credit score, and the label being repayment) and make a prediction about the likelihood that a new loan would be repaid based on the credit score of the individual requesting it (output).

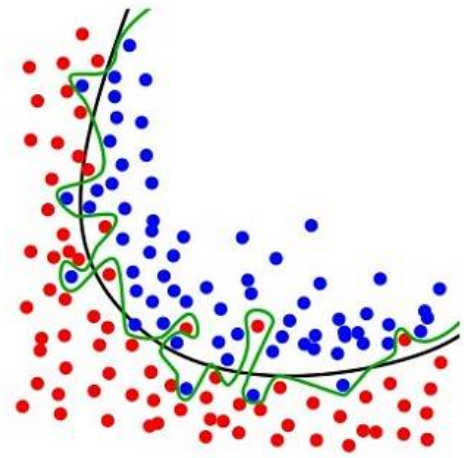
Another method is Unsupervised Machine Learning, which uses input data that does not include labels. In a sense, there is no 'teacher' in this type of machine learning at all. Rather, the programme itself is asked to group data points, identify patterns and discover relationships that would typically be missed by human analyses due to the complexity of the relationships or the size of the dataset.

Tasks

There are two common possible predictive tasks that we could ask our machine learning models to complete: classification tasks and regression-based tasks. A classification task is one in which we are trying to assign new information to a group, for example fraudulent payments versus legitimate ones. In this case, the model will produce an output that assigns a probability score to the likelihood that a new transaction is fraudulent. Beyond answering yes/no-type grouping questions (and the associated probability), we might also be interested in knowing a numerical value, say the expected lifetime value in dollars for different types of users, so we can decide which ones to spend money targeting with a marketing campaign. This would be a regression-type task, and the output would be a numerical figure.

How do the models work?

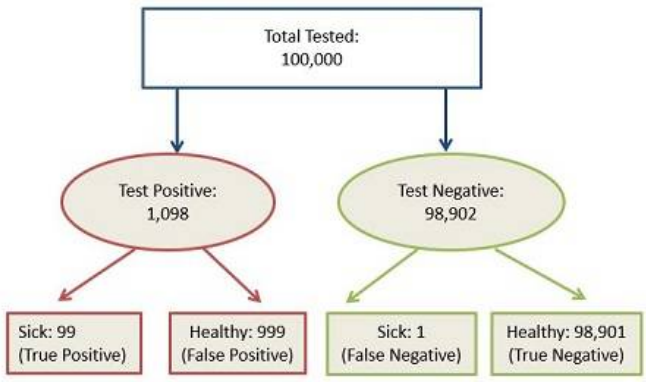
The goal of a machine learning model in this context is to take training data (usually a lot of training data) and to predict whether or not a new observation is likely to end up carrying a certain label. Ideally, our machine learning model has summarised the features of the data in a way that lets it make accurate predictions on new unlabelled data. If not, it will simply memorise the contours of the training data (a problem known as overfitting) and cannot be used to make predictions about new observations. The model should stop 'learning' when it has minimised its predictive error on new observations, not when it has successfully classified every observation within the training data correctly. The image below shows two types of observations (blue could be 'repaid the loan' and red could be 'did not repay the loan'). The green line is an example of overfitting, where the model will have low predictive power on new data, as it has been moulded to fit the training data set precisely. The black line shows a more reasonable classification line; not all observations are accurately captured, but the model is likely to have more success with new unlabelled observations than the model represented by the green line.



Source: EliteDataScience

What does it mean to make a prediction?

Machine learning produces results that rely on probability, and interpreting probabilistic information is something humans do not do well, even experts in a particular field. There is a well-known example from the medical community where doctors are asked to estimate the probability that a patient has a specified disease, given that they have tested positive for that disease. [1] The set-up for this question is that the test is 99 per cent accurate, with one per cent returning either a false positive or a false negative. False positives would be those who do not have the disease returning a positive test result, and false negatives would those who do have the disease failing to test positive for it. Finally, the disease itself is specified to be extremely uncommon, only 0.1 per cent of the population has this disease. So, if you test positive for the disease, what is the probability you have the disease? Many individuals would answer 99 per cent – a figure which *actually* represents the probability that you test positive if you have the disease, quite different from the question posed to doctors. Starting with a sample population of 100,000, we can categorise all the outcomes, as shown in the diagram below.



Starting with 100,000 people, and the fact that the disease impacts 0.1 per cent of the population, 100 people will actually have the disease (100,000 x 0.1 per cent). This means that 99,900 people do not have the disease, however we know that our test is accurate 99 per cent of the time, meaning it is inaccurate one per cent of the time. Combining those two pieces of information we would expect that of the 99,900 people who do not have the disease, 999 (99,900 x one per cent) will still test positive. Similarly, of the 100 people who have the disease, one of them will test negative (100 x one per cent). Putting these figures together, we can find the answer to both questions posed above. The test is 99 per cent accurate given that, of 100 people with the disease, 99 of them will test positive. However, given a positive test result, the likelihood that the person actually has the disease is a little over nine per cent; of 1,098 people who tested positive (99 + 999), only 99 of them actually had the disease (99/1,098 = 9.02 per cent).

When looking at the results from probabilistic modelling, it is important to understand what is actually being presented to avoid common pitfalls as in the example above. There are two parameters that can be targeted within a machine learning model in order to improve the predictive power. The first parameter to target (and minimise) is the instance of false positives; this parameter is

referred to as precision, and is defined as the number of genuine positives over all predicted positives. Precision would correspond to the nine per cent figure arrived at above. The second parameter to target is sensitivity or recall, measured as the ratio of true positives over the actual positive values. In this instance it would be of those sick (100) how many were correctly identified as positive – this is the 99 per cent figure from above (99/100). The results achieved will be highly sensitive to the estimated true occurrence of the disease as well, with high variability in the false positive rate depending on the estimate used. To use the same example from above, if the incidence of the disease is 50 per cent rather than 0.1 per cent, then the two numbers are interchangeable. There will be 50,000 people who are infected and 50,000 who are not. Of those who test positive 99 per cent will be infected (49,500/50,000, with 500 falsely testing positive), and of those who are sick 49,500 will test positive with 500 falsely testing negative, again for 99 per cent.

Precision and recall/sensitivity can be thought of as subcomponents of the model's overall accuracy, and separating the two of them will allow us to assess the impact of the two different types of errors and make tradeoffs, especially for asymmetric misclassification cost/importance, when one misclassification cost is significantly greater than another. In reality, there is a tradeoff between precision and sensitivity/recall, leading to a decision that has to be made as to which is more important. We can move our precision to 100 per cent if we increase the criteria for flagging such that only definitive fraud (for example) is ever flagged, but sensitivity/recall will suffer, since some instances that are fraud will not meet the stringent criteria for flagging them as fraud. This is easiest to understand as a balancing act between your false negative and false positive rate. Focusing only on a low false positive rate (after all, you risk alienating customers or employees by flagging their legitimate activity as fraud) can lead to policies that let fraud occur, allowing more false negatives than the business can afford to bear. A balance is struck between these two, generally involving an additional estimation of the loss given fraud, and the lost revenue from employee/customer attrition if the flag rate is too high.

For fraud professionals, it's relevant to understand what tradeoffs were made in specifying the model, so as not to confuse precision for sensitivity/recall when interpreting the results produced by the model. Just because a model produces a positive result doesn't mean that the activity it's designed to screen for actually occurred. In our example above, medical professionals interpreted the sensitivity/recall (probability you test positive, given you have the disease at 99 per cent) as the same thing as the precision (probability you have the disease, given you test positive at nine per cent).

Further complicating the interpretation of results could be the fact that the true rate of occurrence may be unknown. In the medical example posed, it was given that 0.1 per cent of the population was actually infected with the disease. Estimates of the true occurrence of a certain type of fraud might be unavailable, thus we cannot estimate the precision of the model. The rate of false positives can vary greatly with different assumptions about the true level of occurrence.

Sources of bias or error

Labelling

Some potential biases may be hard to spot unless you go back to the way the training data was labelled. Remember that in a Supervised Learning Model we are relying on human experts to provide labels that facilitate the learning. Sometimes the biases are actually encoded in the way labels are generated. For example, in predicting recidivism data may be encoded as 'committed a crime' when the more appropriate label might be 'was arrested'. The first label already presumes the conclusion, equating getting arrested with having committed a crime, ignoring those who prove their innocence later.

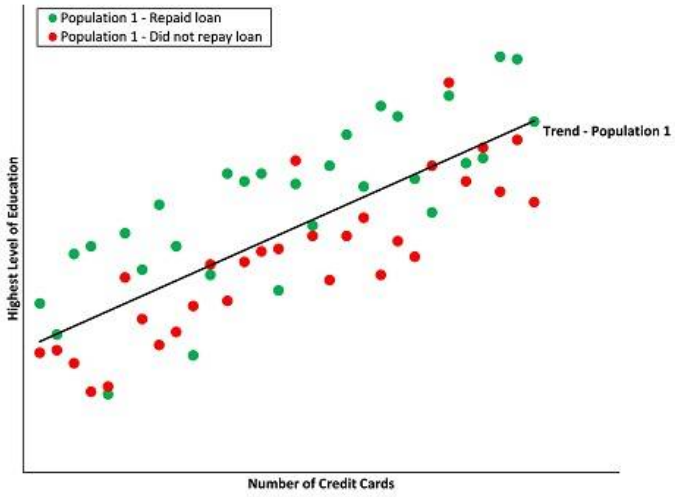
It's also possible that a given dataset was compiled using different standards, but merged and labelled using a single standard. For example, imagine two rental offices, one records renters who pay even one day late as 'late payment', the other gives renters a three-day grace period before recording 'late payment'. When the records from the two offices are merged, no difference between the two types of 'late payment' is noted. As a result, the population renting in the second property may never show 'late payments' and the population renting in the first property will look far more unreliable. As a result, the learning algorithm may find characteristics associated with the first property that it concludes are indicators of future 'late payment', perhaps even identifying characteristics that are specifically different from the second property's renters when, in reality, if property two were held to the same standards as property one it might show an equivalent number of late payments. These types of data issues create problems before the learning model itself has even entered the picture.

Sensitive attributes

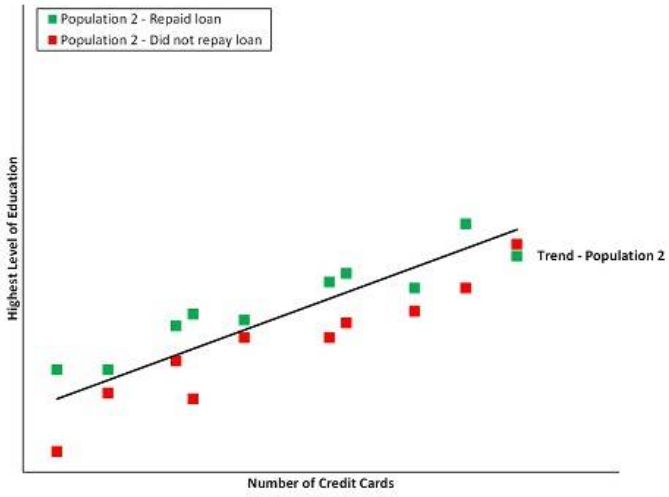
In some instances, a sensitive attribute is removed from analysis, with the good intention of producing 'fair' results. Unfortunately, if that attribute happens to be highly correlated with one or more of the other attributes that are included, removal of the sensitive attribute can actually create more unfairness in the results produced by the model.

For example, consider a lending model in which one of the factors that is known is whether or not the individual is Caucasian or non-Caucasian. For the sake of fairness, the team building the model decides to exclude this as a variable from their analysis. This leaves them with two variables to consider in their simple model – number of credit cards the person has already, and highest level of education achieved. Charting each of the two populations yields the following results – individuals who repaid historical loans are shown in green dots, individuals who did not repay are shown in red. The black line gives a reasonable delineation between the two

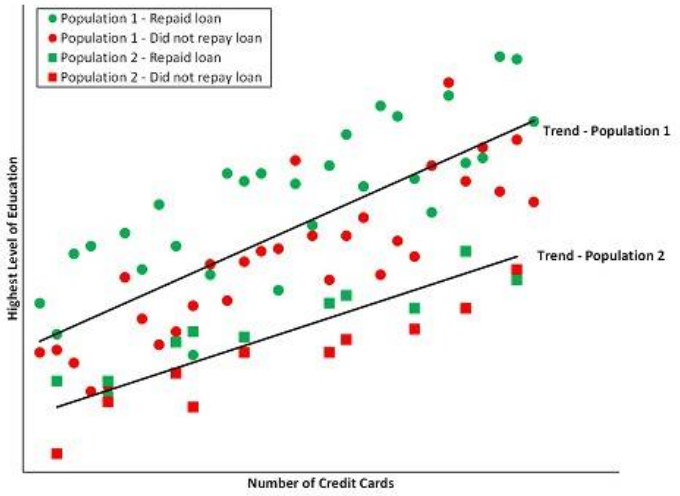
populations, and might be adopted as a lending criteria. The first chart, Population 1, illustrates the data for the Caucasian population.



Shifting to a chart of the non-Caucasian population, we can see that there are fewer data points, but still enough of a trend that it is possible to find a line that represents a good lending cutoff.



But remember, for the sake of fairness, we decided to treat both groups (Caucasian and non-Caucasian) the same, which brings us to the final chart.



In this instance, the line we find that provides decent lending criteria is dramatically unfair to the non-Caucasian population. Non-Caucasians, who, when measured as a group, would be likely to get loans (and repay them) are instead denied loans. Without explicitly accounting for the two different groups, the model will fit to the majority population with more available data points, in this case leading to loan denial for the minority population.

Minority populations

In fact, by definition there is less data available in general for minority populations. Further, many minority populations are unlikely to have access to the kinds of services and devices that would start the data collection process early on in life, the way that many majority population participants are likely to have. So less data may be collected, effectively making these individuals invisible when modelled in conjunction with the majority population. There are many well-known examples of data insufficiency causing problems that appear to be racially driven. Famously, MIT Media Lab researcher Joy Buolamwini and co-author Timnit Gebru tested facial scanning systems that were designed to identify whether a photograph contained a man or a woman. For photos of white men, the systems were accurate 99 per cent of the time. For African-American women, the system worked only between 66 and 80 per cent of the time (depending on the system). [2] It was very clear that the programmes had been trained on photos that were overwhelmingly of white men (the population with the most available data) and could really only successfully identify members of that population.

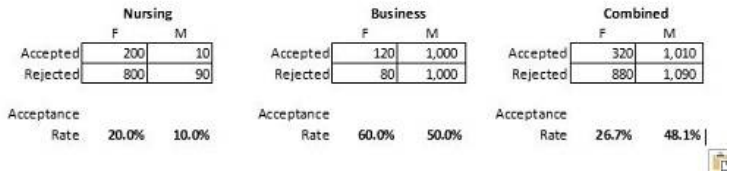
Data affiliated with minority populations, not limited to those that are minority populations by virtue of race, create unique challenges for machine learning applications. Another way to frame the issue would be to consider fraudulent transactions. Let's imagine that fraudulent transactions represent a very small percentage of a given data population, say only one per cent of total observations. A model that was accurate 99 per cent of the time could be constructed purely by virtue of assuming every transaction was legitimate, but we wouldn't necessarily call that a robust prediction model. One technique that can be employed in these situations is to introduce under-sampling of the majority population or over-sampling of the minority population so that the programme has an opportunity to learn genuinely to spot the right type of observations. Generating synthetic samples would be an additional approach.

Hidden variable/unwarranted association

In 1973, UC Berkeley became worried that it would be the first university to be sued for sexual discrimination in its admissions processes. Key to the university's concern was evidence that 44 per cent of male applicants had been admitted to the school, while only 35 per cent of female applicants had been admitted. [3] Upon closer examination it was discovered that there was actually a bias *in favour* of women. How can that be possible? It comes down to the size of the subpopulations that were grouped together, as represented by the departments that men versus women were applying to. From the original paper: "The proportion of women applicants tends to be high in departments that are hard to get into and low in those that are easy to get in to."

So, on a departmental basis, admissions seemed to favour women in that a higher percentage of the women who applied to the department were admitted, when compared to the percentage of men who were admitted to that same department. Because more women applied to difficult-to-enter programmes (as opposed to the less stringent programmes men applied to), in aggregate they fared more poorly in terms of overall percentage admittance. This is an example of Simpson's Paradox, wherein the trend found in the whole population can be different (or opposite) to the trend of its constituent parts. Consider the simple example presented below in which there are only two departments – nursing and business, with nursing as the far more selective programme. As you can see below, more female applicants are admitted to both of the programmes on a per cent basis, but because far more men apply to the

less selective business programme and there are far more male applicants in total, on a combined basis more men are admitted than women.



Nursing	F	M
Accepted	200	10
Rejected	800	90
Acceptance Rate	20.0%	10.0%

Business	F	M
Accepted	120	1,000
Rejected	80	1,000
Acceptance Rate	60.0%	50.0%

Combined	F	M
Accepted	320	1,010
Rejected	880	1,090
Acceptance Rate	26.7%	48.1%

Another famous example is the survival rates on the RMS Titanic, [4] where it was found that survival rates between third class passengers and crew members were roughly the same, at about 24 per cent. This seems impossible to justify narratively, until you take into account the missing variable of male and female within the two groups. Women survived at a much higher rate than men in both groups, but there were relatively few female crew members. So even though the female crew members survived at the highest rate of any group (87 per cent), the effect of that rate was lost due to the far greater number of male crew members who did not survive. Male third class passengers fared the worst (at about 16 per cent), but because there were more female third class passengers than female crew members, the female survival rate helped to push them back in line with the aggregated survival rate of the crew.

These hidden or confounding variables can make data analysis that is done on aggregated statistics difficult to interpret. Conclusions generated about 'fairness' can be exactly wrong in light of Simpson's Paradox, or simply because of poor conception and design from the beginning of a project.

Unintended consequences can be quite serious in light of hidden variables. Consider office supply retailer Staples' decision to adjust the prices of their products for anyone shopping online who lived close to a rival company's physical store. The idea seems rational – offer discount prices to customers who have the option to visit a competitor's physical store nearby; charge those without that option more. What Staples didn't consider was the distribution of physical stores, which tend to be concentrated in higher income areas, rather than lower income ones. The result of the seemingly rational pricing rule was that (on the surface) it looked like Staples was engaged in predatory pricing against lower income individuals and offering discounts to higher earning customers. [5] Staples failed to consider that race, income and other sensitive factors are highly correlated with geographic location.

Feedback loops

Once released from a training environment into full production, a learning algorithm should continue to improve based on its interaction with new information. However, these types of programmes are susceptible to learning errors that are asymmetric in nature. This can create problems later on, when working with the results of these programmes.

Consider an individual who is granted a loan under an automated screening process; since a loan has been extended, the programme will have the opportunity to learn from the outcome of its decision. If the loan is not repaid, it can make adjustments to correct that error later on. A false negative (this person is unlikely to default) will thereby be corrected, but what about false positives, where the programme concludes that an individual is likely to default and rejects the loan?

It is here where the learning becomes asymmetric. While false negatives create new learning, once a candidate for a loan is rejected, no further data is generated related to the outcome that person ultimately arrived at. If an individual is rejected at one company, yet gets multiple loans from another lending company and repays them all, something is wrong with the decision-making criteria within the programme, but the programme never has an opportunity to learn from that mistake. This is especially troublesome in light of all of the other potential problems noted previously. For mislabelled data, covering a minority population, which may be impacted by confounding variables, a programme could easily generate predictions that screen against perfectly qualified candidates and will never have the opportunity to learn from its mistakes, as no information about those false positives would be available for analysis.

Conclusion

Law enforcement is increasingly relying on automated programmes to drive policing – think of the problems associated with facial recognition next time you are presented with an image match – or when given a risk assessment score on the likelihood that an offender will commit another crime. Are there likely to be omitted variables in the programme that produced that result? Companies are also increasingly turning towards automated tools to perform compliance, monitoring and surveillance functions. When you are confronted with evidence produced by one of these programmes, don't allow your judgement to be subsumed by the automated output, especially when deciding if the output is meaningful and does not merely reflect chance correlation between many competing variables. Algorithmic tools can be a valuable addition to decision-making but, in the words of Nobel Prize winning economist Ronald Coase, "If you torture the data long enough, it will confess to anything."

Notes

1. *Calculated Risks: How to Know When Numbers Deceive You*, by Gerd Gigerenzer, Simon & Schuster, 2002, pp 42-48, pp 124-126. See also https://opinionator.blogs.nytimes.com/2010/04/25/chances-are/.
2. www.wired.com/story/photo-algorithms-id-white-men-fineblack-women-not-so-much.
3. www.ncbi.nlm.nih.gov/pubmed/17835295.
4. www.titanicinquiry.org/BOTInq/BOTReport/botRep01.php.
5. www.wsj.com/articles/SB10001424127887323777204578189391813881534

Alex Russell CFA is the director of institutional and complex litigation for Bates Group LLC (www.batesgroup.com), where he manages cases related to institutional disputes involving trust or banking entities, investment banking or sales and trading cases, as well as those involving ultra-high net worth individuals. Alex co-leads Bates' Big Data Analytics segment, with a particular focus on the use of data analytics in market manipulation or fraud cases. He is based in Oregon, United States.

Feb 4 2019