How the Smallest Data tracks the biggest assets

Big Data is a massive misnomer. Many people believe that meaningful insights can only come from vast quantities of data. Indeed, many companies cling to the belief that, to be better informed, they have to collect ever more data. Often, however valuable intelligence can be drawn from the sparsest of data – as this fascinating case study proves. *Wolfgang Emmerich*

Costly and complex assets

Train engines are costly assets controlled by a very complex and devolved network of stakeholders:

- For a start, there are the train owners – typically finance companies – who bank-roll rolling stock…
- They lease the engines and carriages to train operators who hold the franchise to operate particular UK routes. Interestingly, these leases are usually considerably shorter (often 5 years) than the lifespan of a rolling stock – typically 30 – 40 years.
- For convenience, many operators sub-contract the servicing of their leased engines to third party contractors. So, effectively, neither the train's owner nor the operator have any direct control over the care and maintenance of this valuable asset.
- And, to complicate matters further, the trains run tracks and with signaling equipment owned by Network Rail – the same body that also decides Britain's railway schedules. So, the owner does not really know how many miles are covered by any specific engine over any given time.

This paucity of data raises very real concerns. The owners know next to nothing about the state of their assets. They have no way of knowing if operators are taking good care of their rolling stock and are abiding by their contractual terms and conditions of their leases. So, they have no idea whether their depreciation assumptions are wildly optimistic, overly pessimistic or bang on the money.

Ten – or even five years ago – absence of data would have derailed attempts to find answers. But, nowadays, you don't need large volumes of data to do valuable data analytics. Modern machine-learning techniques can provide outstanding insights with relatively small datasets… these techniques can do things that the human brain is no longer capable of performing.

A data Black Hole

The information available for each train engine is minimal. The only real dataset available to owners came from the locomotive's on-board, fuel-tank sensors. Every five minutes, these transmit information on changing fuel levels to a central data store. So, in theory, it was possible to track the levels of diesel consumption for each engine over a measured timeline.

In addition, there is a limited amount of publicly-available information. Network Rail provides a publicly available API that can be used to obtain information on when trains depart from and arrive in particular stations. Our team was able to tap into these sources to establish train movements. We could record precisely when a particular train – typically comprised of two engine units, front and rear, with carriages in-between – arrived and left any given station. Furthermore, this data could be crosschecked for accuracy. Train schedules could be compared with actual departure and arrival information to flag up any delays or incidents.

The snag is, Network Rail do not know which particular engine is operating at any particular time on any specific scheduled service. That is a data black hole – and it gets worse. Train owners aren't interested in how much fuel is left in the tank. They want to know which engines are in use and which are not? Why are some not working (are they broken-down or simply surplus to requirements)? Which particular engines are paired up to form scheduled services on which lines? And precisely where is each engine at any one point in time?

These insights simply aren't available. But in the absence of information, machine-learning suddenly comes into its own.

Machine-learning – the Science of Inference

Our first task was to clean the data because the fuel tank readings were pretty 'noisy'. The read-out graph of the diminishing diesel levels showed big spikes and troughs. These were caused by fuel surges as the train corners, accelerates and brakes.

To remove this noisy time series, we used an L1 Kalman filter. This provided us with a much clearer data reading from which we could unlock valuable insights…

Using a support vector machine, we started to look for recognisable patterns. Focusing on small datasets, we were able to identify time zones. For example, when the fuel level rose quickly within a short period of time it was almost certainly on a stationary refueling stop. Other periods showed the train to be in 'Hotel Mode' – stationary but still using fuel to heat carriages and provide electricity for the train.

Forensic deductions. The analysis then became even more forensic. Our team studied the train's movements over a particular time. From this data it was possible to deduce stationary periods for the train to be recommissioned or refueled… running periods between stops… the frequency of stations… the distance between these stations… and even the probable route and type of service (local, commuter, express etc.).

Engine usage. With the use of standard machine-learning techniques, a very credible and revealing picture of train usage was emerging. Trains with less than 10% movement were deemed to be 'unused'. Engines with under 50% movement were considered to be 'lightly used'. Locomotives with more than 50% movement were classified as 'heavily used.'

Loco 4328 – a case study. Locomotive 4328 is a case in point. By aggregating the patterns which the support vector harvested, we created a coarse-grained visualisation that tracked the train's usage over a 4-week period. For almost the entire period, the engine was unused – a clear indication that it was in the workshop undergoing a fairly major service, repair or upgrade. However, the reassuring news for the concerned owner and operator, was that 4328 returned to 'heavy usage' at the end of this period.

Fuel efficiency. We then turned the spotlight on the critical issue of fuel efficiency. By classifying the data in a different way, the statistics revealed new insights. The stats showed exactly when and where there was high fuel usage (more than half a litre per minute), medium fuel usage and low fuel usage.

Engine matching & intelligence. Then our team used weighted maximum patching algorithms to discover which trains had been paired for different routes. That's when we discovered some closely matching patterns. In some instances, the similarity was so strong, it virtually guaranteed that both engines were paired at the front and back of the same train. By aggregating the data, we could prove that certain engines are routinely paired. Another valuable nugget of operational data.

Mapping journeys. Finally, it is possible to map engine journeys with a high degree of accuracy. By matching the pairings classification to Network Rail's movement data – once again using the weighted matching algorithm
– we could track one train throughout a hardworking day. Leaving Harrogate at 07:34 it travelled to Kings Cross, then on to Newark North Gate, before returning to Kings Cross and finishing the day in Sunderland.

That was quite a journey and it all started with a simple fuel reading!

Take-home messages

Let me close with these thoughts…

Machine-learning techniques, like support vector machines, neural networks, smartmatching, kernel methods and other analytics approaches have become extremely powerful and accessible over the last five years. What's more, we do not have to implement these machine-learning tools ourselves – we can simply re-use open-source libraries.

No single machine-learning technique will deliver all of the answers. You have to know how to combine the right mix of algorithms to derive meaningful insights. This combination is not always obvious or straight-forward. It sometimes requires trials to run it past a domain

expert. For example, the success of this project hinged on a very agile and collaborative relationship between data scientists (with their deep understanding of machine-learning techniques) and the domain experts.

You really don't need large chunks of data to undertake meaningful automated analytics. These fuel time series are literally only a few kilobytes per engine, but they have generated insights that would have been impossible to gain manually. Train asset owners would never be able to derive this level or depth of information without adopting these techniques. Even on very small datasets, machine-learning techniques outstrip and outperform anything that humans can deliver!

Wolfgang Emmerich
CEO UK & Partner

wolfgang.emmerich@zuhlke.com
+44 20 7113 5349