Market Abuse Surveillance TechSprint (July 2024) video Transcript.

Team 9. DHI

Delegate 1

Good afternoon Ken Rodericks, Co founder and CEO of DHI and together with Ross Pearson, Co founder and CTO of DHI. Now we are from Australia, so we brought the warm weather for your weekend, so you're welcome.

And we are a purpose based organisation that uses technology and AI to drive positive change. Now I'll take a moment just to acknowledge the team that has worked on this TechSprint. Our typical approach is to help our clients move from data to knowledge to wisdom by using multimodal AI techniques like Bayesian networks and large language models to find evidence to make decisions.

Now I'm just the opening act and I'm going to transition to the main act over here to take us through the more interesting bits. Now, he's no Taylor Swift, but Ross is equally as good, I promise you.

And Ross is going to take us through how we've approached the Tech Sprint and some of the work that we do for the Australia Securities and Investment Commission in Australia. Ross.

Delegate 2

Thanks, Ken. I'm not sure I can live up to that greeting, but let's see how we go. So Ross Pearson, it's a pleasure to be here. Now I would like to talk to you about two things today.

Firstly, what we did during the Tech Sprint, and that was to use Bayesian networks and anomaly detection to find unknown unknowns or patterns in the data that you would not normally see.

The second thing we did we, sorry, I want to talk about today is what we didn't get a chance to do in the text Sprint and what we'd love to do outside the confines of the sand pit, which was to then reduce the false positives from the anomaly detection by using automated causal explanation.

So a quick prime of everybody.

Firstly, anomaly detection, it's the process of finding unusual or unexpected patterns in the data. The graph to your right just shows you the outliers. That's what we're looking for. Bayesian networks, on the other hand, probabilistic graphical models that express uncertainty. And this is an example of one of the BNS that we pulled together out of the sand pit.

So why did we choose this approach?

Well, firstly we wanted to focus on unknown unknowns, those things that you rarely see or you don't see at all. We wanted a approach that could handle with noisy or missing data and that can deal with uncertainty. And we had a lot of data that we wanted to learn from, and obviously we wanted to be in a position to continue to learn.

So what did we do during the TechSprint?

Well, firstly, we built a Bayesian network. We learnt both the parameters and the structure from the data that was in there. Then we took the data that we'd learnt the model from and we reran it through the model to find anomalies. So we scored every time event. And in this graph here, this sort of describes a particular instrument over time. And the spikes or the dips that you can see in that graph represent the anomalies.

Then we sought to explain what the anomalies actually meant.

So we internally generated a number of visualisations and did analysis. And I'll take you through one of those examples here. So this is the same anomaly chart with volume data superimposed over it. And what you can see here is there's some volume spikes across time. In most cases where you see those blue spikes, you'll see a red anomaly spike coming down. So we we're just seeing an explanation for what that spike is, but not in every case. So some of the spikes don't have explanations that can be attributed to volume, in which case it would be another explanation.

So I want to take you through one example here. So just to ignore that blue line that goes straight up, that's just erroneous data. But the small red circle in the middle is a, a, a trading dip for Kia Group in 2019. So the first thing we wanted to do was explain what caused that dip. So we collected a number of documents and news articles and internal disclosures and quickly identified that this was a surprise profit loss. Then we wanted to actually explain the anomalies associated with that.

So if you see in the other circle with the two dipping anomalies coming down, they're the ones we investigated. What we discovered was that one particular seller, seller X sold a larger, a significantly larger volume on both of those two occasions than they normally would. And this was prior to the release of the announcement that caused the price dip. So well, not saying that this is this this may because for investigation for former market abuse.

So those last steps were quite laborious. It took time to do that investigation.

So the second part of my talk is well, what can we do to automate that? So we start with the signal that red circle, there is a trade event. Now I just want to caveat this. The slides we've got here is the platform that we use for ASIC in Australia. It's a slightly different use case. It's compliance of listing rules. However, the technical approach applies.

So just bear with me, but imagine that this is in fact the same approach that we'd use for market abuse detection. So we start with a signal. We then automate the collection of large news media, any formal disclosures from the company, and then we extract structured data. So another automated step. The reason that we do this is a lot of the data that we collect is in unstructured format, such as PDFs and web pages. And a lot of that structured data we can use for our models.

In the last step, we then use a combination of artificial, sorry, intelligent agents and large language models to generate a causal explanation. We do this by collecting all the data that we we're pulling all the data that we collect and we apply a set of rules and a process called retrieval augmented generation to mitigate hallucinations and we generate a contextual explanation. So the idea here is that through this process of court creating a explanation, we can categorise those explanations and then prioritise a lot of the anomalies.

So in conclusion, I've shown you Bayesian networks for anomaly detection, I've shown you a process of applying large language models to generate an automated causal explanation, and combined the idea is that you could then rapidly prioritise your market abuse investigations.

And that's it for me.

Thank you.