Market Abuse Surveillance TechSprint (July 2024) video Transcript.

Team 8. Kaizen

Delegate 1

Welcome to the Kaizen TechSprint presentation. My name is Charles, head of Operations for Surveillance. In the crowd we've got Simon, market Abuse and Surveillance Director, and I am just from Quintana.

Delegate 2

I am data scientist at Kaizen.

Delegate 1

So in this project, the FCA asked us how AI could be used to revolutionise market abuse surveillance by incorporating anomaly detection. In this presentation, we aim to show the results of our findings. To encourage discussion, we took a small set of variables, used a range of innovative and appropriate measures and then visualise the results to help explain and verify them.

We will show a selection of features that we believe are indicative of market abuse and we will then show 3 examples that show the algorithm is working as expected. In the future, this technology will be built into our product to initiate investigations. The variables are split into three categories based on how we treated them, price movements, user and instrument profiles and statistical and geometric measures. We'll go through each of those in turn. Those variables then drive 2 anomaly detection systems before we visualise the results.

Delegate 2

Thank you. So as just have said, the first technique that we use is neural networks. This is testing and detect price movements that are inconsistent with normal market volatility. So we try to detect here is prices that could have some driving factor outside the normal volatility. So this is how the model interprets what's happening.

The grey area here is the part of the data that the algorithm uses to learn how the price behaves. And then as you can see, the red dots there mean that the algorithm thought there was something a bit different about the pricing there. And then maybe some driving factors that are different volatility. So transactions in the vicinity of this movement could be seen as potential insider trading, for example.

The next set that we created was done by matrix factorization. So the aim here is to take users, traders and instruments, the instruments that they trade and create profiles so that we know this user should be using or it's likely to use this instrument, this instrument and this other instrument.

The hypothesis is that a user or a trader that makes profits out of an instrument that is unlikely or out of the ballpark for that type of user could be engaged in a lot of market abuse. And these are slides that explains how we did this profiling, but there's no time, so we can please skip for the next one. Yes.

And the third large group of variables we created was done with more conventional methods like standard deviations, normalisation, etcetera. For example, we created one variable that that studies how a user or trader transacts volumes across time for a certain instrument. And if a volume is unusually high for them and for other people in the group, then that could be seen as suspicious.

So this is what we have done.

These many variables that we can discuss in the Q&A, the these distributions, how they should look like set for the next stage, which is anomaly detection. We see on the left hand side towards the 0 are all the behaviours that seem very normal, like usual price movements or a person transacting an instrument that they normally do are. Towards the right we see a very small group of anomalies, things that are unusual in the group and to the person. Now we run two different anomaly detection systems. One is isolation forest because it's quite fast, but it has got vulnerabilities and we run another one as well as a factor to make up for the weakness of, of, of the algorithm.

So we have got many variables we studied, but we also only use three here for visualisation. So in one axis we've got trade volume, another one changing transaction pattern, another one is returns. It's just a measure that tells us that price moves in wide swings in sometimes in the vicinity of of money of abuse. The shader, the deeper shade of Reds shows where anomalies are flagged, as you can see, are more towards extremes of the axis. And that shows that those people or those transactions are more likely to be abuse.

This illustration of how our anomaly addiction works, but we can also leave it because we don't have so much time.

Delegate 1

So in this example, we're looking at the highest suspicion score and as is often the case with this type of algorithm, initially this has returned a data error.

So the price move is so extreme it was prioritised over other variables. And in addition, it's very rare for for this instrument to have been traded. So we need more discussion and refinement so we can better handle this kind of exception going forward.

The lower suspicion score has a very indistinct price move and it's also very likely that the trader trades this instrument and a very high proportion of traders in general have traded this instrument in the middle of the distribution. As you might expect, price movers average likelihood to trade the instrument is average and the number of trades is consistent with what other traders are doing.

Delegate 2

Thank you. And because the time frame we were given to study the data was quite small, we're aiming to show that algorithms are working consistently with good practise.

So most of the suspicious scores, as you can see, are below nought .6. That means that most transactions were found to be quite regular, and very few separated other ones with high anomaly scores.

Delegate 1

So right now we're looking at all variables as equally important, but that's probably not the case. We need to look at more variables to grain to gain greater insight. It should also be noted we're only looking at GBI syns and natural persons due to processing power and time constraints.

In the future, we need to look at more and more meaningful variables, assess their weightings and refine the algorithm. Ideally we could then look at more granular time frames of pricing and look at technologies further. Technologies such as Internet searches, databases and large language models.