FRAUD DETECTION AND PREVENTION: A DATA ANALYTICS APPROACH

BY SESHIKA FERNANDO
TECHNICAL LEAD, WSO2
# TABLE OF CONTENTS

1. Fraud: The Bad and the Ugly.................................................................03
2. A New Opportunity for Fraud Detection...........................................03
3. Creating Fraud Rules from Domain Expertise...............................04
4. Addressing the ‘False Positive Trap’ with Scoring..........................05
5. The Next Steps - Learn, Adapt, Repeat............................................06
   5.1 Markov Modelling.................................................................07
   5.2 Machine Learning...............................................................08
6. Building a Powerful and Comprehensive Fraud Detection System....09
7. Enterprise Fraud Detection Using the WSO2 Analytics Platform.....11
8. Conclusion......................................................................................12
9. References.......................................................................................13
1. FRAUD: THE BAD AND THE UGLY

Fraud has become a trillion dollar business today. The Association of Certified Fraud Executives (ACFE) has reported that businesses lose around 5% of revenue to fraud every year. And when that ratio is applied to the gross world product (~ $80 trillion), it translates to a glaring loss of $4 trillion a year to fraud.

Today, fraud is prevalent across almost all industries and has become more rampant and complicated. Enterprises are constantly struggling to implement effective and efficient fraud detection systems. The use of traditional fraud management systems for this purpose, however, has proved to be ineffective as fraudulent activity itself is becoming a cross-channel affair.

A few years ago, businesses found it sufficient to run nightly batch fraud detection processes that utilize some complex statistical models managed by a few data scientists. Yet, with an increase in transactional channels (online, mobile, etc.) and the shift towards real-time decision making, there is a pressing need for real-time fraud detection solutions that are able to detect patterns over multiple channels and are able to self-learn and update itself. This will eliminate the need to maintain groups of highly skilled, but expensive data scientists that are economically prohibitive for most businesses (sparing only the largest banks and financial institutions) in order to be protected from fraud.

2. A NEW OPPORTUNITY FOR FRAUD DETECTION

Fraud rules and statistical models alone are no longer sufficient to detect fraud in real time within this complex landscape. The ability to combine batch analytics, streaming analytics and predictive analytics with domain expertise (Figure 1) is imperative to set up an effective fraud detection system. Furthermore, these analytics should be able to model both known and unknown forms of fraudulent or anomalous activity.
3. CREATING FRAUD RULES FROM DOMAIN EXPERTISE

With businesses of all types, forms, and sizes being vulnerable to fraud, the very definition of fraud itself is open to interpretation. For example, what might be perceived as possible fraudulent behavior in a small online store might be business as usual in a large multinational organization. Therefore, it is important to allow businesses to define what they perceive as fraud and hence enable them to convert expert knowledge in their domain into a set of fraud rules. All transactions, individually and collectively, will then be compared against these fraud rules in real time and get flagged as fraud once a rule is violated.

The following examples show how easily domain knowledge can be converted to complex event processing rules:

**ABNORMAL TRANSACTIONAL QUANTITIES**

One of the basic checks any business needs to have is a rule against abnormal transaction quantities. The threshold need not be defined by a human being, but can be derived from past and present transactional data itself. To illustrate, the first query given below will calculate and maintain the hourly moving average of quantities for each item. The second query will check incoming transaction quantities against a dynamic threshold (99th percentile -> average + 3 standard deviations) that reflects current trends, seasonal changes, etc., since the threshold itself gets updated for every incoming transaction.
from TxnStream#window.time(60 min)
select itemNo, avg(qty) as avg, stdev(qty) as stdev
  group by itemNo
update AvgTbl as a
  on itemNo == a.itemNo;

from TxnStream[itemNo==a.itemNo and qty > (a.avg + 3*a.stdev) in AvgTbl as a]
select *
insert into FraudStream;

**TRANSACTION VELOCITY**

Businesses can evaluate how frequently subsequent transactions are done in an item and set a threshold to filter out large deviations. This can be set at a global level or at a per customer level based on historical trading patterns. If it is abnormal for a customer to do three or more transactions within 5 minutes, a complex event processing query can be easily written to capture this pattern in real time.

```java
from e1 = TransactionStream ->
e2 = transactionStream[e1.cardnum == e2.cardnum] <2:>
within 5 min
select e1.cardNo,  e1.txnID,  e2[0].txnID,  e2[1].txnID
insert into FraudStream
```

The above query will send out an alert when two or more transactions follow an initial transaction from the same card number within 5 minutes. The alert will include information about each transaction. In the above query, it provides the card number and the transaction IDs of the first three transactions (that are required for this rule to be violated).

Similarly, all domain, product or customer specific anomaly knowledge can be easily captured using generic complex event processing rules.

**4. ADDRESSING THE ‘FALSE POSITIVE TRAP’ WITH SCORING**

With the ability to write rules to capture all domain specific fraud logic, there’s also the danger of losing customers to overprotective fraud rules. Businesses are often stuck in a Catch-22. Enterprises, for instance, would need to be careful to not block or inconvenience a customer who’s genuinely trying to purchase the most expensive item in a store, but at the same time have the necessary controls to prevent a fraudster from attempting to make an expensive purchase using a suspicious IP address.

Scoring is a simple mechanism that can be used to address these challenges. It enables enterprises to use a combination of rules (as opposed to a large number of individual rules) with a weight attached to each rule, and generates a single number that reflects how well a transaction performed against multiple fraud indicators.
The weight attached to each rule will correspond to the severity of the rule. For example, a rule written to capture transactions done by a blacklisted IP address will be assigned a bigger weight than a rule that captures transaction quantities larger than the 99th percentile.

Assuming that you’ve written 7 individual rules that capture various fraudulent behavior patterns, you can write a scoring function to generate a single number using all 7 rules as follows.

\[
\text{Score} = 0.001 \times \text{itemPrice} + 0.1 \times \text{itemQuantity} + 2.5 \times \text{isFreeEmail} + 3 \times \text{highTransactionVelocity} + 5 \times \text{riskyCountry} + 5 \times \text{suspiciousUsername} + 8 \times \text{suspiciousIPRange}
\]

The rule names above are self-explanatory and hence rationalize the difference in the weights attached to each of them.

5. THE NEXT STEPS - LEARN, ADAPT, REPEAT

The sections above described the process of converting domain expertise about fraudulent behavior into real-time fraud rules. It also explained the use of ‘scoring’ to reduce false positives, and hence be notified in the event of only a fraudulent activity.

Although this will safeguard an enterprise against an amateur fraudster or an individual who’s well versed with stealing account credentials and using them to carry out fraudulent activity, your business is still not completely safe. Most career fraudsters who belong to organized crime groups usually know about generic rules and scoring functions. They are able to mobilize fraud strategies using their international criminal network and elude the rules and scoring functions that most fraud detection systems contain.

Predictive analytics techniques like Markov chains and machine learning can be used to counter these threats by handing over the definition of ‘abnormal’ or ‘fraudulent’ behavior to an algorithm, which will self-learn and self-adjust according to both legitimate and fraudulent activity patterns in real time.
5.1 MARKOV MODELLING

Markov models are stochastic models used to model randomly changing systems where it is assumed that future states depend only on the present state and not on the sequence of events that preceded it. When building a Markov model for transactional fraud detection, we employ a three-step process – State Classification, Probability Calculation, and Metric Comparison. We apply these steps to historical data, in order to build the model, but let the model update itself be based on real-time data.

STATE CLASSIFICATION

As the first step, we classify each event based on certain qualities of the event. For example, assuming the events are credit card transactions, we can classify each transaction based on

- Amount spent – low, normal or high
- Whether the transaction includes a high priced item – normal or high
- Time elapsed since the last transaction – low, normal or high

Based on the above sample classification, we would be left with 18 (3x2x3) possible states.

PROBABILITY CALCULATION

Now that each transaction is classified into a particular state, we count all the state transitions in the historical, sequenced data, and produce state transition probabilities. This will result in a state transition matrix as illustrated in Figure 2 (only a subset).

<table>
<thead>
<tr>
<th></th>
<th>LNL</th>
<th>LNH</th>
<th>LNS</th>
<th>LHL</th>
<th>HHL</th>
<th>HHS</th>
<th>HNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>LNL</td>
<td>0.143510</td>
<td>0.002152</td>
<td>0.207060</td>
<td>0.095459</td>
<td>0.007166</td>
<td>0.509172</td>
<td>0.035481</td>
</tr>
<tr>
<td>LNH</td>
<td>0.136484</td>
<td>0.409425</td>
<td>0.188628</td>
<td>0.051126</td>
<td>0.113801</td>
<td>0.000711</td>
<td>0.099825</td>
</tr>
<tr>
<td>LNS</td>
<td>0.074190</td>
<td>0.082927</td>
<td>0.401801</td>
<td>0.156532</td>
<td>0.012045</td>
<td>0.201713</td>
<td>0.070792</td>
</tr>
<tr>
<td>LHL</td>
<td>0.052885</td>
<td>0.134071</td>
<td>0.328956</td>
<td>0.286087</td>
<td>0.106753</td>
<td>0.063064</td>
<td>0.028184</td>
</tr>
<tr>
<td>HHL</td>
<td>0.316206</td>
<td>0.055719</td>
<td>0.051524</td>
<td>0.069597</td>
<td>0.062291</td>
<td>0.444638</td>
<td>0.000025</td>
</tr>
<tr>
<td>HHS</td>
<td>0.204606</td>
<td>0.002722</td>
<td>0.043194</td>
<td>0.159342</td>
<td>0.238314</td>
<td>0.006382</td>
<td>0.345440</td>
</tr>
<tr>
<td>HNS</td>
<td>0.007737</td>
<td>0.371359</td>
<td>0.326846</td>
<td>0.082564</td>
<td>0.021326</td>
<td>0.015835</td>
<td>0.174333</td>
</tr>
</tbody>
</table>

Figure 2
METRIC COMPARISON

As a final step, when we receive transaction sequences in real time, we compute one or more metrics based on the transition probabilities for that event sequence and compare that against a threshold to flag rare event sequences as probable fraud. Miss Rate metric, Miss Probability metric and Entropy Reduction metric are some of the metrics we use for this purpose.

And it doesn't just stop there; we classify every real-time event and update the probability matrix that allows the model to learn, based on trends and seasonal behaviors in both genuine and fraudulent activity.

5.2 MACHINE LEARNING

We also employ machine learning algorithms to model the ‘normal’ behavior of events and hence detect deviations from the modelled ‘normal’ behavior in real time data. This falls under the unsupervised learning category where we give the algorithm a bunch of data and request it to tell us which ones are normal and those that are not. Unfortunately, due to the large disparity in the ratio between the number of fraudulent data to the number of legitimate data, most classification algorithms do not work well for anomaly detection problems. Therefore, we use ‘clustering’ mechanisms, which allow the modelling of ‘normal’ behavior as clusters, and anomalies (fraud) as deviations from those clusters.

Figure 3
As illustrated in Figure 3, the clustering algorithm will define a number of clusters that hold normal or legitimate events. These will be represented as a bunch of cluster centroids and densities. Any incoming event situated outside of any of the existing clusters (measured by centroids + densities) will be flagged as anomalous (fraudulent).

A fraudster can even review this piece of information, but would still not be able to bypass these detection mechanisms as it’s defined using a combination of historical and real-time data; and it’s self-learning and adjusting according to both legitimate and fraudulent activity trends.

**6. BUILDING A POWERFUL AND COMPREHENSIVE FRAUD DETECTION SYSTEM**

A combination of the above methods can be used to create a very powerful fraud detection system. However, once a possible fraudulent event is detected, an enterprise must dig deeper to understand whether there are any other events/relationships that this event is linked to. Therefore, it pays to query other events that have something in common with the event that was flagged as possible fraud.

In order to do this, you must first make sure that you store all incoming events (not just fraudulent ones). Then when we get a fraudulent event, you can query that event store for other related events.

Once you query related events, this can provide useful visualizations and further querying facilities for a fraud analyst to discover possible relationships within the events. This will enable the discovery of large fraud rings and collusions.

In the following examples, Figures 4 and 5 show geographical relationships and intensity diagrams based on other transactions, related to the fraudulent transaction, that have been queried by the fraud analyst.
7. ENTERPRISE FRAUD DETECTION USING THE WSO2 ANALYTICS PLATFORM

The WSO2 Analytics platform can be used to create an efficient and effective fraud detection solution that encompasses all of the above methods. WSO2 Data Analytics Server (WSO2 DAS) (as explained in Figure 6) comes with batch, real-time, predictive, and interactive analytics capabilities that can be used to convert domain knowledge into generic rules, implement fraud scoring, utilize Markov models and data clustering to model unknown types of fraud and obtain interactive data querying and visualizations.

The Generic Rules, Fraud Scoring and Markov Modelling is implemented through the real-time analytics engine of WSO2 DAS, while the persisted data can be used by Predictive Analytics to create Clusters of normal data that will be fed back into the real-time analytics as operators. Furthermore, the Interactive Analytics capabilities will be used to provide meaningful visualizations and further querying to the Fraud Analyst who would want to dig deeper using the alerts that were generated by the real-time rules.

As illustrated in Figure 6, we can create a comprehensive enterprise fraud detection system using the different analytics capabilities of WSO2 DAS.

The WSO2 Analytics platform combines into one integrated platform real-time and batch analysis of data with predictive analytics via machine learning to support the multiple demands of Internet of Things solutions, as well as mobile and web apps. It replaces WSO2 Business Activity Monitor 2.5 and also has the capabilities to organize and analyze data that would have been previously inaccessible or unusable. Moreover, it builds on the fast performance of the open source Siddhi CEP engine developed by WSO2 by...
adding streaming regression and anomaly detection operators to facilitate fraud and error detection.

As part of WSO2’s analytics platform, WSO2 DAS introduces an industry first with the ability to analyze both data in motion and data at rest from the same software. The platform provides a single solution that enables developers to build systems and applications that collect and analyze information and communicate the results. It has been designed to treat millions of events per second, and is therefore capable to handle the volumes big data and Internet of Things project. WSO2 DAS 3.0 will be released in Q4 2015. Refer to wso2.com/products for further details.

8. CONCLUSION

Even though fraud has become a trillion-dollar business today, the effort to distinguish a fraudulent activity from a legitimate one can be compared to looking for a needle in a haystack. Every enterprise knows the importance of adopting a comprehensive fraud detection system that’s effective in detecting and preventing fraudulent activity and, at the same time, be efficient enough to remove legitimate activity by a valued customer. It’s a fine balancing act that would eventually help enterprises to address these challenges, which in turn will aid them to prevent revenue losses as a result of fraud.

The way forward is to convert domain knowledge regarding fraudulent behavior to real time fraud rules, use Markov modelling and machine learning to detect unknown abnormal behavior, and use scoring functions to reduce the number of false alarms being raised. WSO2 DAS offers interactive analysis capabilities that enable fraud analysts to further crack down on larger fraud rings and organized criminal activity by using the useful visualizations and intuitive query capabilities of WSO2’s analytics platform.
9. REFERENCES

1. True Cost of Fraud 2014

2. Stop Billions in Fraud Losses using Machine Learning
   https://www.forrester.com/Stop+Billions+In+Fraud+Losses+With+Machine+Learning/fulltext/-/E-res120912

3. Big Data In Fraud Management: Variety Leads To Value And Improved Customer Experience
   https://www.forrester.com/Big+Data+In+Fraud+Management+Variety+Leads+To+Value+And+Improved+Customer+Experience/fulltext/-/E-RES103841

   https://www.forrester.com/Predictions+2015+Identity+Management+Fraud+Management+And+Cybersecurity+Converge/fulltext/-/E-RES120014

5. Markov Modelling for Fraud Detection
   https://pkghosh.wordpress.com/2013/10/21/real-time-fraud-detection-with-sequence-ining/
ABOUT THE AUTHOR

Seshika Fernando
Technical Lead
WSO2

Seshika’s current area of focus is in the application of WSO2 products in financial markets. She started her career at Millennium Information Technologies and later joined Commercial Bank of Ceylon on completion of her postgraduate studies. She holds a BSc (Hons) in Computer Science from the University of Colombo, Sri Lanka, and an MSc in Finance from the London School of Economics, UK.

She is a British Chevening Scholar and holds a national patent and is also a visiting lecturer at the Institute of Chartered Accountants Sri Lanka, where she lectures for the MBA program of University of Southern Queensland, Australia.

ABOUT WSO2

WSO2 is the only company that provides a completely integrated enterprise application platform for enabling a business to build and connect APIs, applications, web services, iPaaS, PaaS, software as a service, and legacy connections without having to write code; using big data and mobile; and fostering reuse through a social enterprise store. Only with WSO2 can enterprises use a family of governed secure solutions built on the same code base to extend their ecosystems across the cloud and on mobile devices to employees, customers, and partners in anyway they like. Hundreds of leading enterprise customers across every sector—health, financial, retail, logistics, manufacturing, travel, technology, telecom, and more—in every region of the world rely on WSO2’s award-winning, 100% open source platform for their mission-critical applications. To learn more, visit http://wso2.com or check out the WSO2 community on the WSO2 Blog, Twitter, LinkedIn, and Facebook.

Check out more WSO2 White Papers and WSO2 Case Studies.

For more information about WSO2 products and services, please visit http://wso2.com or email bizdev@wso2.com