Busting Financial Crime with TIBCO
— Ana Costa e Silva, PhD
Senior Data Scientist, TIBCO Software

What if you could use just one financial crime fighting solution that would empower your business users to improve handling of financial crimes such as anti-money laundering (AML), credit card fraud, trade surveillance, or medical fraud? Current financial crime fighting systems have two disadvantages:

• They flag too many false positives that cause investigators to focus on the wrong cases.
• They involve manual procedures that result in investigations taking too long to complete.

TIBCO’s solution addresses both these disadvantages.

PRODUCING MORE RELEVANT ALERTS
TIBCO’s approach to fighting financial crime places machine learning in the center of the crime detection system. Machine learning models use historic data to learn how to spot risky or abnormal behavior exhibited by transactions, clients, suppliers, or other players. It uses two types of models:

  Supervised learning algorithms, which tell us how similar to past fraud a new transaction is.
  Unsupervised learning algorithms, which tell us how odd a new transaction seems when compared to past transactions.

The first model guarantees accuracy, the second the ability to adapt to changing realities.

These models learn from history and then these learnings are applied to the present, either in real time or in batch by simply scoring current data against the models. Transactions that are found to be risky or odd beyond a certain threshold will be manually investigated. The setting of the threshold is a business decision supported by what-if analyses in TIBCO Spotfire®.
The following examples explain how the system can be trained to detect different types of crime.

**SUPERVISED LEARNING ALGORITHMS**

Modeling with supervised learning algorithms involves obtaining data of confirmed fraudulent and non-fraudulent cases. For example, for a list of transactions monitored for AML, one column contains a value of “1” when past transactions were fraudulent and a value of “0” when non-fraudulent. Decision trees, random forests, neural networks, support vector machines, and logistic regression are all examples of supervised learning algorithms. Given all else known about each transaction, they generate optimal ways (models) of separating the 1s and 0s to obtain true positives with minimum incidence of false positives.

A model is a summary of a pattern in the historic data and therefore a much smaller representation of the original data. For example, the decision tree in Figure 1 could have been learned from millions of lines of historic data.

If value of the transaction > 40K then

If proportion of value over average balance is >85% then

Probability of a transaction being AML = 80%, else

Probability of a transaction being AML = 30%

else, Probability of a transaction being AML = 10%.

Figure 1 Example of the result of a decision tree algorithm on a big data set.

TIBCO Spotfire is a visual analytics software tool that allows you to run advanced statistical models. An easy-to-use Spotfire template can guide the business end-user through the steps of building and testing model types, even if the user has no deep knowledge of statistics or data science. The user just needs to understand the business.

Figure 2 below shows an example of such a template. On the left hand side under “Response” the user chooses the variable he wants to model. As an example for credit card fraud, the variable would contain the value 0 for all non-fraudulent transactions and 1 for the fraudulent ones. Under “Predictors,” the user chooses which features, or columns, of the data he wants to use to build the model. Then under “Models,” the user has the option to try different types of models.

![Image of a Spotfire template where business users can train and test different supervised models.](image-url)
Spotfire can automatically detect the presence of new columns in a dataset, and add them to the list of Predictors, which means the user can quickly and dynamically adapt a Spotfire template to include new features. The user is also free to include a number of algorithms or just one. When the user presses the button “Fit models in parallel,” Spotfire calls out to its statistics engine to run the relevant calculations. Results, including any quality tests performed, are published back to Spotfire. In the example in Figure 2, the user only needs to know that the best model is arguably the one with highest Area Under the Curve (ACU) as shown in the table.

The Spotfire data function that does all this work in the background typically needs to be developed by a data scientist. Data functions are calculations using your preferred statistical scripting language or workflow tool and are designed for collaboration. Once the data function is created, it can be easily shared. Any analyst can use data functions without needing to know any coding. All appropriate business users are empowered to make better decisions, including creating fraud models, without being exposed to unnecessary complexity.

Spotfire supports data functions in different statistical engines, such as TIBCO Enterprise Runtime for R (TERR), which is embedded in Spotfire, as well as in open-source R, SAS, Matlab, KNIME, and Lavastorm.

**UNSUPERVISED LEARNING ALGORITHMS**

Using only supervised models is not enough. In fact, some companies may be starting a financial crime fighting unit without historic knowledge about which transactions were fraudulent and which were not. Even when you do have historic knowledge, it is never certain that all past fraud cases were correctly identified. And even when this certainty is fairly high, fraudsters are creative, and if one strategy did not work, they will try a new one that will leave a different fraud trail behind.

TIBCO recommends using a combination of supervised and unsupervised models, and the TIBCO solution accommodates this.

Unsupervised models have no requirements for prior knowledge about which transactions were fraudulent and which were not. Without a goal variable per se, this type of algorithm aims to capture what is “normal” in the data and which different types of normal there are. Clustering algorithms and self-organizing maps are examples of this type of model. When applied to financial crime data, these methods allow profiling normal operations and spotting unusual ones. Unusual does not mean criminal, it means warranting human verification.

Figure 3 below contains an example of a Spotfire template that, with minimal training, a business user can use to develop an unsupervised model. In this case, a well-established matrix operation called Principal Component Analysis (PCA) is used to represent all transactions. The relevant components are shown on the axes, where normal transactions appear close to the origin of the chart (0,0 point) and abnormal ones farther from that point. The distance of any new transaction to this origin is a measure of its oddity. This relevant information is not the result of any human assumption, but is derived directly from the pattern drawn by the whole history of transactions. Transactions that are odd beyond an agreed threshold should be investigated.
GOOD FEATURES MAKE GOOD MODELS

Any predictive model is as good as the features put into it. One challenge companies often face is identifying which characteristics to focus on to identify fraudulent events. Good fraud features are those that allow spotting unusual behavior. Often external business consultants are hired to suggest such features, but it’s been our experience that the best features are already intuitively known by the experts in the firm. Consultants often gather business knowledge from internal experts and translate it into quantifiable features that can be extracted from their databases using SQL. For example, in AML, some relevant features are:

- Total amount of cash withdrawn. Unusually high values warrant an investigation.
- Value of withdrawal as a proportion of the account owner’s average balance. High value withdrawals in relation to the average account balance are worth checking.
- The amount of time between the withdrawal and a previous deposit of a similar amount.
- The value of the withdrawal as a proportion of the value of the previous deposit.
- The value of the withdrawal as a proportion of the mean withdrawal of customers who share similar characteristics (gender, age, income, etc.) or of companies in the same economic sector, size, and region.
- Whether the account owner has a family relationship with one of the bank staff, etc.

Many of these features are gathered and monitored with systems such as Actimize, but are treated with individual, and not mathematically optimized thresholds. Mathematical models can combine all features optimally.

A different set of features would be used for finding fraud in medical insurance claims. Imagine an insurance policy that covers 100% of all emergency claims. An obvious fraud for this policy is to declare routine procedures as emergency. The incentive for this potential fraudulent behavior is specific to the set-up of this particular insurance policy. This is why your business people, who understand the terms of your different policies, will know best what would be odd and potentially fraudulent behavior. A conversation with the database administrators is all they need, to together, derive the optimal SQL that helps grasp the relevant features.
Relevant features for this policy might be:

- Total number and proportion of emergencies by doctor / clinic / patient
- Time between an emergency appointment and the purchase of the prescribed medicine
- Time between emergencies per patient and per family

When features have been crystallized into SQL, Spotfire can collect this data straight from the relevant database and visually portray how transactions behave. For example, on the left hand chart of Figure 4, it’s easy to spot large withdrawals. This seems to be a useful thing to keep track of, especially if the money withdrawn represents a large portion of the average balance. In Spotfire you can select the people who satisfy each or both conditions and list them. A quick investigation of a few cases will provide a better feel for the usefulness of the features for detecting the criminal activity.

If historic data already contains the information on which transactions were fraudulent or non-fraudulent, this knowledge can inform the search for fraud revealing features. On the right hand chart of Figure 4, a zoomed-in box plot shows that transactions with higher value have also been more likely to be fraudulent.

Network charts provide another rich source of insight to identify people who have a big impact on the network as a whole, for example, organizations receiving a large amount of cash deposits from many different people.

Figure 4 Visual analysis of AML related features with no past knowledge about which were fraudulent. The left visualization shows total cash value withdrawn per transaction. The right, a distribution of selected variable by status showing that higher value transactions are more often fraudulent.

One does not necessarily need to visualize every feature, however. Especially for big data containing more features, it may be impossible to visualize them one by one. The results of the supervised model can guide the search for the most relevant features for spotting fraudulent activity. We should then visualize these individually. Figure 5 shows which features have a higher contribution to the model (the longer the bar, the more important the feature). Although Figure 4 showed that “Value” (eighth in rank) is important, other features are more powerful at distinguishing fraudulent transactions.
A combination of visualizing the ranking of the features as well as the detail of the individual features is important for a number of reasons:

1 Validation of the model's quality. Maybe your best feature is so good because it is part of the answer and should therefore be excluded. For example, if you inadvertently included the total value of fraud that was stopped as a predictor, that will obviously (and erroneously) appear as the best predictor.

2 Correlation is not causation. It is necessary to ask questions that lead to a better understanding of the reality being predicted.

3 Validation of the data's quality. Were you expecting a different feature to have more power than what is showing? Perhaps there are data quality issues causing a lack of relevance, or maybe outliers introduced a bias. These quality issues can be quickly spotted in a visualization.

4 Surprising top features. Sometimes predictors expected to be irrelevant turn out to have huge predictive ability. This knowledge, when shared with the business, will inevitably lead to better decisions.

5 Inspiration for new features. Sometimes the most informative features are the reason to delve into new related information as a source of other rich features.

6 Computational efficiency. Features with very low predictive power should be removed from the model as long as the prediction accuracy on the test dataset stays high. This ensures a more lightweight model with a higher degree of freedom, better interpretability, and potentially faster calculations when applying it to current data, in batch or real time.

SELF-LEARNING ABILITIES
Another benefit of ranking features is that, in time, organizations become better at spotting financial crime. First, by gathering better features; second, as the system generates alerts, transactions are investigated and classified as either a true or false fraud alert. This knowledge is fed back into the system and informs the next round of the supervised model. So naturally, the more often the system runs (with the latest information available), the better it gets at finding future fraud that looks like past fraud. This is what the term “machine-learning” actually means, a machine that is capable of learning on its own.

HOW TIBCO DEPLOYS MODELS IN REAL TIME
Because a model is a summary of historic data, it can be as light as an equation and live beyond the data. Once the user is satisfied with the quality of the model, a press of a button is all that’s required to send that model to the real-time event processing engine that will monitor transactions as they occur. TERR, TIBCO’s statistics engine, has excellent integration with all of TIBCO’s event processing products to support this automated capability.
REDUCING INVESTIGATION TIME
How does TIBCO propose dealing with the second biggest flaw in current fraud detection systems: the fact that each alert takes too long to investigate? TIBCO’s event processing software warns investigators in real time as soon as a potentially fraudulent or odd transaction is spotted. There are many options for how this warning can occur, for example by SMS, by email, or simply by accumulating risky transactions in a pre-specified location.

When an alert is produced, a pre-defined Spotfire investigative template is automatically loaded with the potential fraudulent data and made available to the investigator. The template will contain all the relevant context regarding the specific transaction—all its intervening events, and data from any relevant data source. The interactive nature of Spotfire will enable the user to complete a swift but solid investigation.

Figure 6A provides an example of an automatically generated Spotfire email report. It includes the ID of the transaction, a link to the respective investigative template, and a visualization providing context. In this case, it shows how the transaction compares against transactions of similar individuals. When the investigator presses the link, a new Spotfire instance, shown in Figure 6B, opens in any choice of web browser for computer or mobile device. This investigative template contains all the relevant context of the transaction. In this case the network chart shows that the sender has had transactions with two more people who had received AML alerts in the past.

TIBCO’s Business Process Management (BPM) system can further ensure that this entire process is auditable by anyone at any time. TIBCO ActiveMatrix® BPM Spotfire® will make sure that all relevant documents are generated, the right people are informed, and the required signatures are collected.
CONCLUSION

TIBCO proposes ONE modular financial crime fighting solution for anti-money laundering (AML), credit card fraud, trade surveillance, medical fraud, and other financial crime, which:

- Monitors ALL transactions in one auditable, repeatable, and self-learning process
- Increases customer satisfaction because it limits involvement to only those potentially exposed to real risk
- Increases investigative team productivity by only calling their attention to risky transactions that can be investigated with instantly provided context for optimal decision-making
- Puts advanced mathematics at your fingertips
- Without requiring a degree in advanced math or computer science, enables real domain experts to apply their knowledge of specific business operations
- Provides all this via easy-to-use dashboards built for business users

Learn more about TIBCO Spotfire, TERR, and other TIBCO products for big data processing and fraud detection by contacting us or visiting www.tibco.com and spotfire.tibco.com.

Figure 6B Investigative template that can be sent to an investigator in real time containing the entire context of a risky transaction.