# Fighting Fraud with Advanced Analytics

**Troy Hiltbrand** is chief digital officer at Kyäni and an associate editor of the *Business Intelligence Journal*.
thiltbrand@kyanicorp.com

**Troy Hiltbrand**

## Abstract

**Business fraud is on the rise and businesses need to have confidence that the person on the other end of the transaction is legitimate and that the transaction is valid. With a large volume of transactions happening in real time, businesses have to quickly identify which are legitimate and which are fraudulent. Advanced analytics provides the platform and method to accomplish this.**

**Advanced analytics provides a mechanism to effectively automate the discovery of transactions that don't quite "look right." Uncovering such fraud can save companies millions of dollars in unrecoverable sales.**

## Introduction

From increased globalization and stricter regulation to a consumer community hungry for personalized digital interaction, doing business today is a challenge. As the business environment grows ever more complex, so does the increase in digital fraud. Not only are consumers becoming more digitally literate and interacting with businesses in entirely new ways, so are criminals looking to identify weaknesses in corporate defenses. In the digital arena, this nameless and faceless threat can make companies feel helpless, but this is not the case. Advanced analytics can provide businesses with the weapons they need to combat this growing problem.

With identity theft and credit card harvesting becoming more common, there is a corresponding increase in the use of these stolen identities and payment methods against businesses. Criminals look for ways to monetize the information they have acquired—opportunities where it can be turned into cold, hard cash—and businesses are often left liable for the costs.

There are two main costs that businesses have to cover. First, the true owner of the breached financial

account has to be refunded the money that was illegally compromised. Financial institutions are good about acting as advocates for their customers in this regard. Second, however, the lost inventory and freight costs of the transaction also have to be covered. Often, both of these costs land squarely in the court of the business that accepted the fraudulent transaction.

Businesses need to have confidence that the person on the other end of the transaction is legitimate and that the transaction is valid. With a large volume of transactions happening in real time, businesses have to quickly identify which are legitimate and which are fraudulent. Advanced analytics provides the platform and method to accomplish this.

These methods and techniques can be highly effective when applied to fraud detection, but they also can be applied more generally to other business problems where the unknown has to be predicted with a high degree of probability.

### Understand the Business

Deploying advanced analytics for fraud management requires that the business fully understands the business situation and deploys the set of analytics that is optimized to address their specific challenges.

The first step in this process is to understand what the perpetrators of fraud are doing. This will be highly dependent on the business and its policies and procedures, including returns, exchanges, and payment options.

Those committing fraud have an agenda. Most often, this has to do with using information to generate financial rewards for themselves at the expense of others. Another common cause of fraud against a business is a perpetrator's personal conflict with the business. This could be based on the business's stance towards a social or political issue, or based on a soured relationship with the perpetrator.

Once you understand the end game of those committing fraud, you can start to look at historical transactions and identify where fraud has happened. If it has happened before, it will probably happen again. The patterns of historical transactions form a basis for anticipating future fraud. Criminals don't like to change their modus operandi unless they have to, so the past is a pretty good indicator of the future.

Unfortunately, with financial transactions, fraud often manifests itself only weeks or months after the transaction. Therefore, a set of relevant data must include many months of transactions.

Unfortunately, with financial transactions, fraud often manifests itself only weeks or months after the transaction. This is when the person whose identity was compromised comes forward to dispute the transaction with his or her financial institution, which usually ends in a chargeback to the business supplying the good or service. Therefore, pulling together a set of relevant data must include many months of transactions so that the patterns of historical fraud can be adequately defined.

### Define a Training Data Set

To learn from the past and predict the future, you need data to develop and train an analytics model. Once you have a business understanding of how a criminal might be using transactions against the company and examples of where it has happened, you are ready to start processing the data.

As you analyze fraudulent transactions, it is essential to develop a profile for each one. This will include information related to the specific transaction and patterns of how a particular transaction relates to others.

During this step, your company will often discover that certain data that would be highly beneficial to include

in the profile is not readily available. For example, with credit card transactions, perpetrators often test out the card by making a few small transactions with it at different businesses before making a large transaction. Raw information about where the card has been used prior to the transaction being investigated exists, but is often outside the purview of what your business controls.

In this case, a surrogate attribute might be required in the transaction profile that can represent suspicious activity that exists outside of the business. Many financial institutions have a process for prescreening transactions. As part of the prescreening process, these financial processors will return data in the form of a flag or score that may be used as this surrogate attribute in the transaction profile.

Depending on the completeness of the data environment and the organizational barriers that exist (whether technical or political), pulling together an accurate profile of fraudulent transactions can be extremely difficult, often requiring time and collaboration across organizational divisions. Some of the data that is most relevant to identifying fraud is sensitive in nature and must be carefully governed to ensure that its use doesn't expose the company to further liability.

Once you finish profiling these transactions, you will have a large list of potential attributes associated with them. At this point in the model development process, you do not have to define the rules associated with determining if a transaction fraudulent or not; you are just arranging the data so that it can be used to develop a statistical model later.

This step might generate hundreds of attributes associated with each transaction. Some of them will have no bearing on defining the model. Others will only be applicable when used in context with another attribute or set of attributes. Still others will be highly correlated to whether the transaction is fraudulent. During the data preparation stage, the goal is to identify a good set of relevant attributes, centralize these attributes, and cleanse the data so that it is reliable for modeling.

Given the costs associated with covering both sides of a fraudulent transaction, even a small amount of fraud can be very expensive for the business. However, examples of true fraud are usually dwarfed in a sea of valid transactions, so it can be challenging to get a good sample size of fraudulent transactions with which to develop your model. At times, this will require that the training set be disproportionately loaded with fraudulent transaction records as compared to a full data set. This will allow the model to be sensitive enough to be able to identify fraudulent transactions in the future.

As you dissect the problem statement and the business environment, your list of factors may include:

- Transaction amount
- Time of day and day of the week when the transaction was made
- Historical quantity and quality of transactions from the same or similar customers
- The true location of the transaction's origin, obtained from attributes embedded in the network packets that created it, not necessarily the information entered by the ordering party
- Frequency of attempted transactions

These will make up the transaction profile.

Within the data set, there is one critical attribute—the indicator of whether the transaction is fraudulent or not. As this is not always available in the transactional system, this will often have to be engineered using a combination of other data sets. Typical sources of data used in constructing the fraud indicator include financial records related to chargebacks, returns and metadata related to the reason for the return, and customer notes. With data such as customer notes, the content is often unstructured and requires preprocessing so that it can be used in the model development process. The resulting fraud attribute needs to have only two potential values: TRUE or FALSE.

In advanced analytic modeling, this fraud indicator is known as the *target attribute*. A target attribute is one that is known during the modeling process, but unknown during the process of prediction. With fraud, the goal is to develop a model that will ascertain this fraud attribute from a set of known attributes.

To enable real-time detection, the profile attributes need to be limited to only those pieces of information that are known at the time of the transaction.

One word of caution: as you identify attributes associated with fraud, ensure each attribute is one that would have existed at the time of the transaction.

Take an example where returns are part of the fraud pattern. Returns only happen after the transaction. If these are part of the model, real-time fraud detection will be impossible because the model will be optimized to take a longer-term position and will only be able to identify fraudulent transactions once the full picture—including returns—is known. To enable real-time detection, the profile attributes need to be limited to only those pieces of information that are known at the time of the transaction.

## Building the Model

The term *advanced analytics* is often misrepresented as a single method or technology. In reality, it represents a category of approaches and not a specific technology. Advanced analytics includes multiple algorithms and algorithmic approaches that use mathematics and statistics to extract unknown information from a known set of data.

When using advanced analytics to detect fraud, the goal is naturally to determine if a transaction is fraudulent. This could be in real time as the transaction is happening or in near time prior to the completion of the transaction and the delivery of goods or services to the perpetrator. When fraud is headed off before the transaction is complete, the costs of lost inventory can be avoided. Additionally, declining the transaction can often eliminate the future chargeback on the account and its associated fees and penalties.

Fraud detection most often falls into a category of analytics known as *supervised learning*. Supervised learning techniques create a model by iteratively cycling over the data, optimizing the model's performance by adjusting its parameters.

With fraud, the model's target is to identify likely fraudulent transactions without throwing too many false positives. True positives caught in a timely fashion can save the company money, but false positives can deter legitimate business activity and negatively affect relationships with valid customers.

As a result of this supervised learning process, you create a process by which a known set of inputs is transformed into a prediction of the transaction's fraud attribute.

In the field of supervised learning, the analytics community has identified multiple techniques for accomplishing this model development and optimization. Depending on the technique, the resulting models range from being easy to visualize and explain to highly complex, where the business has to treat the model as a "black box" and accept understanding the inputs and the output, but not the transformational model.

Methods such as decision trees are easy to lay out graphically, which makes it easy to walk people through the process from raw inputs to fraud decision. The decision tree uses a set of divisions in the data to lead to a final answer. The decision tree can be simple, only dividing the data into a couple of sets, or it can be multiple levels deep—using different attributes to subdivide the data into smaller and smaller groups, each representing a defined output. As the target variable is either TRUE or FALSE, the decision tree algorithm will have multiple routes to each potential answer.

Other methods, such as neural networks and support vector machines, are more complex and so more difficult to track back from a result to the raw data inputs.

Ensemble methods use multiple techniques jointly to transform raw inputs into a fraud decision. To fully understand the result, you must understand both the ensemble scoring method as a whole and each of the methods used to generate input for it.

Some analytics techniques only function with certain types of data. For instance, methods using vector mathematics and linear algebra to develop the model will often require that all of the attributes be numeric. Such methods would require that attributes that are not numeric either be converted from categorical information into numeric values or be left out of the model development process.

As these techniques and their resulting models are based on known statistical and mathematical concepts, even these complex algorithms can be built from scratch with every part of the "black box" understood. The challenge with this approach is that it can be a costly investment for a company whose goal is to simply develop an effective model that will allow it to predict whether transactions are fraudulent in a timely fashion. This is where you have to understand the relationship between the cost of developing a model and the value it provides the business by heading off fraudulent activity.

The fastest way to develop these fraud models is not by coding the model generation from scratch but through the utilization of a platform which has prebuilt tools for performing this function. Leaders in this space include SAS, RapidMiner, IBM, Microsoft, and Oracle. Another popular tool for developing fraud models is R, an open source platform developed and maintained by a community of analytics practitioners. There are companies, such as Revolution Analytics, that provide commercial support for the R language.

R and Python, another open source programming language, have together become very popular among the advanced analytics community for developing models. It is not always as easy to program models in R and Python as it is with the commercial platforms, but the entry cost for the technology is much more attractive.

The fastest way to develop these fraud models is not by coding the model from scratch but through the utilization of a platform which has prebuilt tools for performing this function.

As you select a tool for developing fraud models, there are a couple of key considerations that need to be part of the tool selection process.

- Different technology platforms used in the model building process will expose differing levels of insight into what the model is doing.

  Some will generate code that can be implemented in multiple computer languages and systems. This code can be broken apart and analyzed to understand what is happening. Others provide only key attributes of the model and the rest remains entirely inside the tool. You will need to determine how much transparency you will require into the model being run.

- It is critical to identify a tool your organization either has or can develop skills with to build the model.

  Effectively acquiring and preparing data, developing a model and deploying it in production are all job functions of what is now known as a data scientist. This role is highly sought after by organizations due to the potential impact of advanced analytic models in optimizing business. This also creates a relative scarcity in the market for individuals who possess these skills. Due to this scarcity of skills in the market, it is important to factor your existing and targeted knowledge of both the tool and process into tool selection.

## Testing for Accuracy

The magic in advanced analytics comes not through simply developing a model, but iteratively testing that model and refining it. During this process, alterations to the data used to train the model and optimization of the parameters used to configure it drive the model to be continually better. Success is achieved when the model is able to accurately identify true fraud, while minimizing false positives.

When measuring the model's effectiveness, it is important to test using data that was not part of the data used in the development phase. During the development process, there is a tendency to over fit the training data. As a result, the model can produce spectacular results when run against the same data that was used to train the model, but will fail miserably when assessing real world data.

## Deploying the Model into Production

Once the business process has been analyzed, a profile created for existing fraud cases, and a model developed, tested, and refined, the next step is to roll out the model into a production environment and allow it to process real transactions.

Oftentimes, practitioners believe that once a model is tested and deployed, it will run forever and continue to return the same level of quality results. The problem with fraud is that as it is uncovered and prevented, perpetrators evolve their tactics, and existing models will be ineffective against these new methods of fraud.

There is no predefined amount of time that a fraud model will run effectively because it depends on how quickly the market evolves as the model exposes and stops fraudulent behavior from happening. In cases where fraud is highly lucrative, models might have to be rebuilt on short time frames such as weeks or months. In less lucrative cases of fraud, the business might be able to go for quarters or years before having to rebuild their models.

The best indicator of a model's effectiveness is to track how many cases of fraud continue to pass through the established filters. As the fraud starts to increase to unacceptable levels, it is an indicator that the models need to be recreated to better address the fraudulent behavior in the market.

In this complex business environment, companies are faced with more challenges as they strive to succeed. Among these challenges, fraud has become more prevalent and thus a more expensive part of doing business. Being able to adequately address fraud can help a business succeed in these perilous times. If done correctly, advanced analytics provides the platform for effectively detecting fraudulent transactions and stopping them before they cost the business its future viability. ■