Fraud Detection

# Fraud?

# Industry frauds Examples?

Fraud Analytics In [The Health Care Industry](http://www.dataiku.com/solutions/industries/healthcare/)

* *Challenges*: Increased cost of providing insurance benefits to employees, genuine data compromised, patient exploitation, higher premiums and out-of-pocket expenses for consumers
* *Data*: Wide variety, from electronic health records to accounting data
* *Typical Use Cases*: Patient billing analysis to determine claim fabrications, financial records and behavioral analysis to determine identity theft, surgical procedure analysis to determine unnecessary services

Fraud Analytics In [The Insurance Industry](http://www.dataiku.com/solutions/industries/insurance/)

* *Challenges*: Ability to obtain accurate data, cost vs benefit of fraud investigation process
* *Data*: Insurance claims, interaction data (social media, cell phone, ATM usage), financial records, police records, hospital records
* *Typical Use Cases*: Insurance claim analysis to predict potential fraudulent claims, injury data analysis to predict likelihood of insurance policy fraud, public sector analysis of fraudulent claim data (e.g., arson, property, auto) to predict areas prone to fraudulent activity (e.g., for premium determination/adjustments)

# Use case?

Credit Card Fraud

* Location
* Items you buy
* Frequency
* Amount

**Feature Selection**

* Transactional Statistics
* Regional
* Merchant Type
* Time based Amount Statistics
* Time based number of transactions statistics

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4770922/>

<http://www.businessinsurance.org/10-most-common-types-of-insurance-fraud/>

<https://medium.com/anomaly-detection-with-python-and-r/how-to-create-good-features-in-fraud-detection-de6562f249ef>

<https://mapr.com/blog/predicting-loan-credit-risk-using-apache-spark-machine-learning-random-forests/>

<https://mapr.com/blog/real-time-credit-card-fraud-detection-apache-spark-and-event-streaming/>

<https://www.semanticscholar.org/paper/Detecting-Credit-Card-Fraud-Using-Periodic-Features-Bahnsen-Aouada/4e5a1f1bf9c7f7f1cc0a8121863ed6e3e233d42a/figure/6>

<https://arxiv.org/ftp/arxiv/papers/1510/1510.07165.pdf>



Fraud Detection Use Cases: Industry-Specific Challenges

Fraud Analytics In [The Financial Securities Industry](http://www.dataiku.com/solutions/industries/banking/)

* *Challenges*: Determination of data accuracy, loss of investor confidence, compromise of corporate reputation, financial loss for company and its investors
* *Data*: Publicly available financial records, corporate records, banking transactions and transfers
* *Typical Use Cases*: Security price analysis to determine likelihood of authentic vs inflated valuations, corporate financial records analysis (assets, liabilities, costs, revenue) to predict securities fraud

References:

<https://www.dataiku.com/solutions/use-cases/fraud-detection/>

Fraud in insurance examples:

<http://www.businessinsurance.org/10-most-common-types-of-insurance-fraud/>

<https://medium.com/anomaly-detection-with-python-and-r/2-features-for-healthcare-fraud-waste-and-abuse-7c262ac59859>

<https://www.berdonllp.com/combating-fraud-in-the-hospitality-industry/>

<https://legaldictionary.net/bank-fraud/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4770922/>

<https://medium.com/@Dataman.ai>

* Detecting fraudulent mobile-phone calls in telecommunications scenarios.
* Identifying fraudulent credit card transactions for banking institutions.
* Identifying fraudulent purchases in retail or e-commerce scenarios.

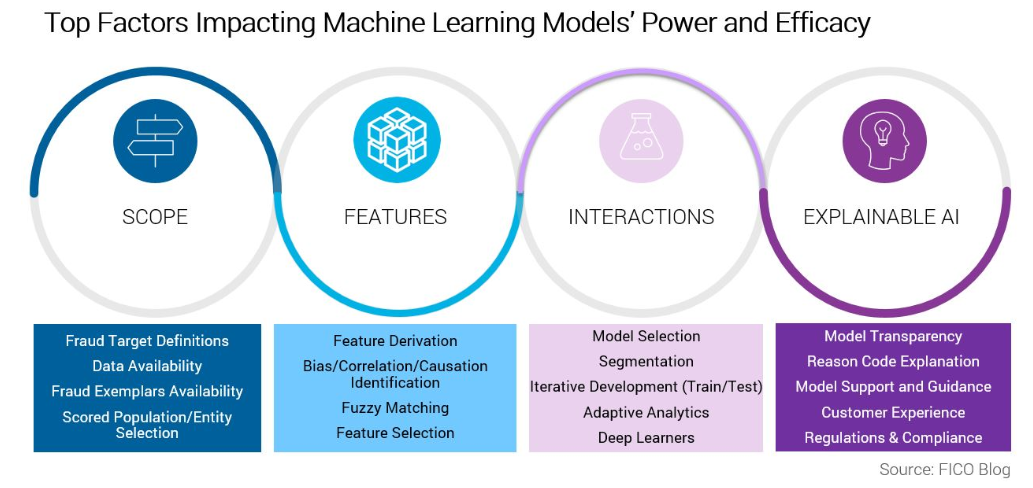
<https://towardsdatascience.com/detecting-financial-fraud-using-machine-learning-three-ways-of-winning-the-war-against-imbalanced-a03f8815cce9>

<https://arxiv.org/pdf/1811.02196.pdf>

<https://www.datascience.com/blog/fraud-detection-with-tensorflow>

<https://dhrubajitdas44.blogspot.com/2017/11/machine-learning-project-3-credit-card.html>

<https://rpubs.com/slazien/fraud_detection>

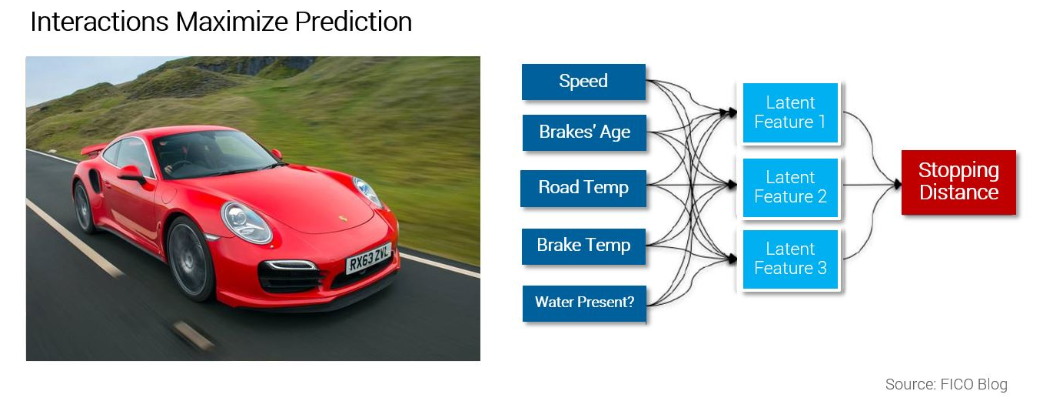


**Scope**

Scope requires setting a reasonable objective for a model, including identifying the target you want to make a prediction about and ensuring there is sufficient data. While a highly focused model only addresses a small, specialized part of a fraud problem, we are able to optimize and refine the model to do that specific job very well. On the other hand, a general model can address a larger class of a fraud types, with the trade-off that it may not perform as well as a specialized model.  Often the best is a combination of focused and general scored models that can be efficiently handled by segmentation via sub-models.

To strike the right balance, we must ask:

* What entity are we scoring — the application, the applicant, the customer, the account?
* What models are built for specific fraud schemes (e.g., synthetic identities, bust-outs) or is the fraud definition more generic (e.g., third-party and first-party fraud) requiring broader defined fraud detection models?
* What historical data is required to support the development of the model types — are there sufficient fraud tags?
* Is the model consortium utilizing various clients’ definitions of fraud types that must be fused to take advantage of broader insight and performance gained from consortium approaches?
* How long is the model expected to be used in production? What is the anticipated lifetime of the mode? Do we need adaptive and self-learning technologies to keep the model robust?
* What are the transparency requirements of the model?
* **Features**
* Machine learning models rely on features derived from the historical data to infer the patterns and relationships within the data. In other words, data doesn’t cut it and we need to create feature detectors of both fraud and non-fraud supported both by the available historical data and the anticipated production data. We need to appropriately define features based on the current historical fraud and non-fraud behaviors, and create features that are generic and that will adapt to new patterns in the production data as fraudsters change their tactics.
* There are many types of features that FICO uses in our own application fraud model characteristic libraries. One example is the use of Network Analytics (SNA), as was discussed in a recent FICO webinar, [Layered Defenses in the Fight Against Application Fraud](http://www.fico.com/en/latest-thinking/on-demand-webinar/layered-defenses-in-the-fight-against-application-fraud). Other examples include behavioral profiling techniques, archetypes clustering, bureau characteristics, output of fuzzy matching, and many more that are a topic for another day.
* **Interactions**
* From feature creation we have a number of feature detectors, each with varying degrees of predictive power, but they must be combined into a score. Machine learning allows the signals in data to be combined in optimal ways to maximize detection and minimize false positives. The relationships learned are far too complex for human analysts, or indeed the data scientist, to impute. The machine learning process facilitates learning of nonlinear combinations of features — latent features — where machine learning explores which combinations of features lead to interactions between signals that strengthen predictiveness.
* For example, consider the raw data inputs needed to determine the stopping distance of a car. You could capture a variety of data points such as ambient temperature, the presence of water, and weight of the vehicle. There are relationships between these inputs that are far more powerful than a simpler model that does not model interaction terms would learn from historical data.
* However, to achieve the most accurate results, the model should also look at interactions between these variables. For example, the relationship between water and temperature indicates the presence of ice — clearly a predictive feature that strongly influences stopping distance.



**Explainable AI**

One of the biggest challenges in operationalizing machine learning in fraud detection is the need for transparency in a highly regulated credit lending environment. FICO is on the frontier of [explainable AI](http://www.fico.com/en/blogs/analytics-optimization/explainable-ai-breaks-out-of-the-black-box/)(XAI), researching and developing techniques to enable humans to understand how the machine learning model reached its score, increase the shelf life of analytic models, and providing real-world applicability for organizations to utilize the power of advanced analytics while meeting governance and regulatory requirements on ML.

References:

<https://www.fico.com/blogs/analytics-optimization/preventing-application-fraud-with-machine-learning-and-ai-part-2/>