Fraud Analytics Using Machine-learning & Engineering on "Big Data" (FAME) for Telecom

Fraud Analytics Using Machine Learning & Engineering on Big Data (FAME) was one of its kind work done in Telecom Fraud Detection. The project was most exciting as it touched upon many unknown territories of data visualization, feature engineering, predictive model building and fine tuning, implementation on Big Data. It challenged us at every step moving forward making the whole journey exciting and overcoming each obstacle a memorable experience. We would like to thanks Indian School of Business and Tata Communications Ltd to give us this exciting opportunity, and also thank to our guide Prof ***Prof Thriyambakam Krishnan*** for his most valuable mentoring and guidance .

1 Project Description

Motive behind telecom fraud, in most cases, is to generate monetary benefits illegally for fraudsters themselves. Globally $46.3 Billion revenue loss due to fraud, out of $10.76 Billion are lost in International Revenue Share Fraud (IRSF).

Table 1-1 Brief on International Revenue Sharing Fraud

How does it happen?	Who are impacted?
• Fraudster creates premium rate numbers in high priced destinations such as Somalia, Latvia, etc. And advertises to encourage people to call-in these numbers. Most of the time, a payment agreement is arranged to create traffic to these numbers. • People with intension to gain easy money access telephony system illegally (e.g. by PBX hacking, sim cloning, stolen sim) to generate traffic to these premium rate numbers and keep the connection alive for longer duration.	• Customers – huge bill to retail and corporate customers due to fraud. Typically deny to pay. • Telecom operators – International Telecom Agreement needs the operator to pay the next operator in chain of the network. So destination telecom operator is benefitted as they get their share of revenue, operators at source and transit carriers at risk of denial of payment and legal hassles.

A timely detection is the only way to minimize this loss.

The problem lies in the fact that the pattern of fraud changes over time with fraudsters adopting newer methods, historical pattern detection mining mechanisms are not efficient. And also volume of data (in order of a few hundred GBs in a day) makes it difficult to process data in timely manner. This paper presents an industrialized solution approach with self adaptive machine learning and application of big data technologies to address this. The approach has been demonstrated for detection of *International Revenue Sharing Fraud* (IRSF) with less than <5% false positive results. However, this approach

is extendible for detection of most of telecom frauds and frauds of similar nature. More than 1.2 TB of history Call Detail Record (CDR) data from one of the top overseas transit carrier has been used to demonstrate this approach successfully.

2 Overall Approach

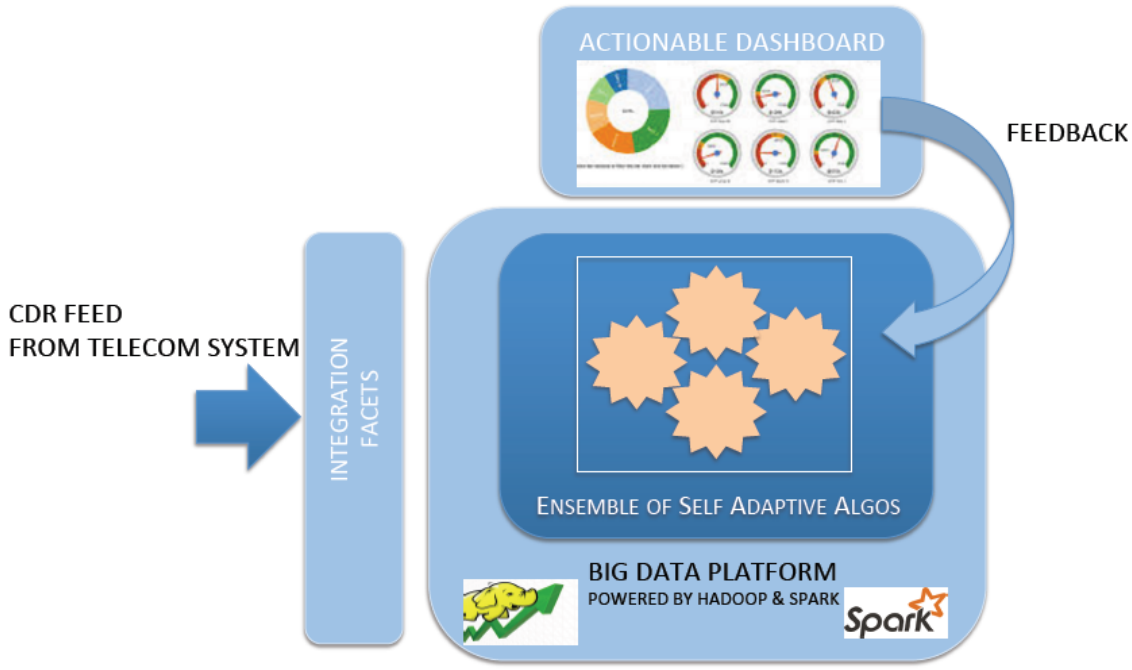


Figure 2-1 Overview of FAME for Telecom

Basic components of FAME are:

- Self adaptive Machine learning methodology
- Actionable dash board for operations and investigations team to act upon the alerts and feedback sent to machine learning model for adjusting weights.
- High performance big data platform for data processing and machine learning

2.1 Machine Learning Methodology

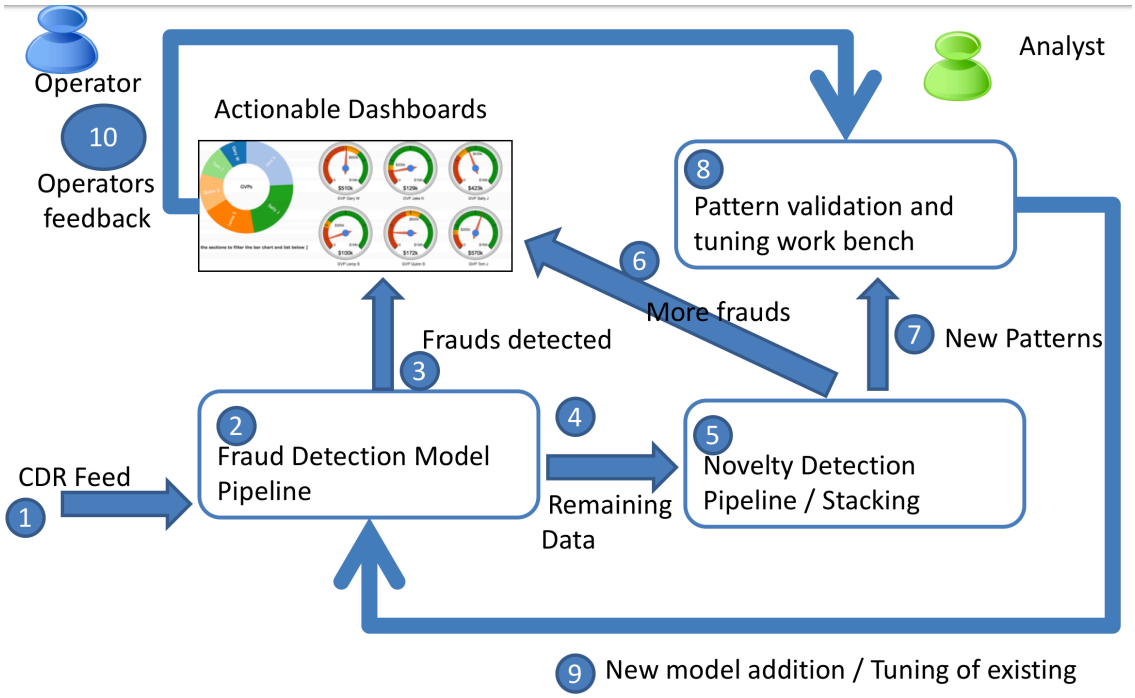


Figure 2-2 Machine Learning Methodology

Machine learning methodology for the approach has 3 different components:

Novelty Detection / Pattern Discovery Pipeline/Stacking

Contextual Anomaly Detection [3] is used to detect novelty in individual algorithms. In contextual anomaly, data points may be normal in other scenarios, but in presence of values of other attribute it is unusual.

Various algorithms have been used to detect anomaly based on type of the anomaly we want to detect, some of them are:

- Outlier based on similarity/distance matrix
- Multivariate probability distribution of normal behavior
- Empirical distribution based on historical profiling of attributes

Pattern validation and model tuning workflow

Discovered patterns are reported through an actionable dashboard and workbench. Based on feedback on the dashboard, appropriate fraud data is used to either build fresh model or tune existing models.

Tuned models are tested against the historical data and patterns before they are pushed to fraud detection model pipeline.

Patterns as well as detected fraud source or destinations are stored and used for smart filtering of future detection.

Detection Model Pipeline

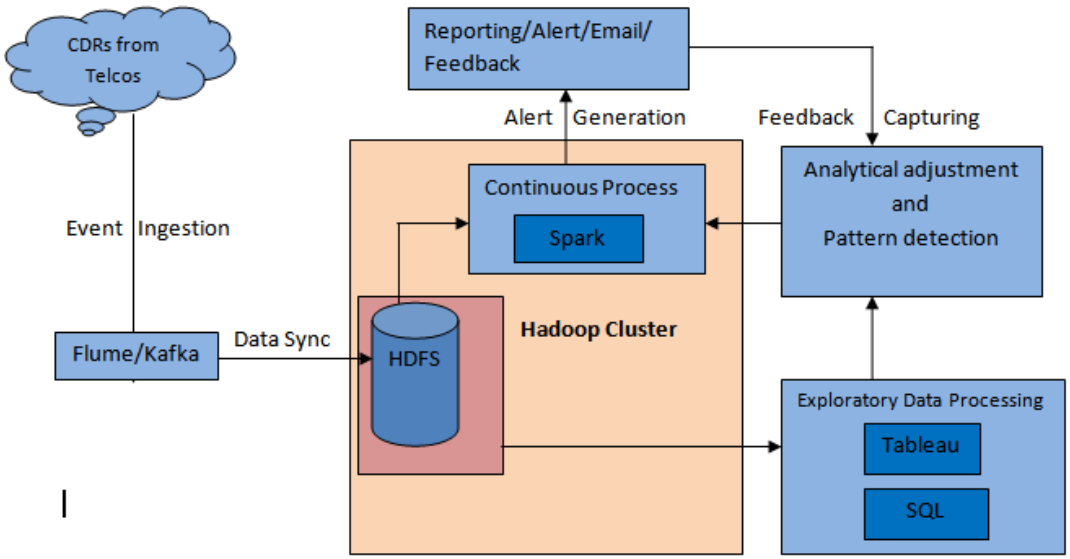
Detection models are created based on fraud patterns detected. Pipeline combines the output of each components through a tunable logic to provide final output.

2.2 Actionable Dashboard

For the fraud detection system to be adaptive, it is needed that feedback goes to the model on the fraud alerts generated and novel patterns discovered. System can be auto-tuned or tuned under human intervention on case-to-case basis. Actionable Dashboard facilitates this along with serving as alert and pattern reporting mechanism.

2.3 Big Data Platform

Big Data platform are leveraged to process CDR in streaming or batch mode and run machine learning algorithms to produce alerts in time efficient and cost effective manner. Combination of Apache Hadoop and Apache Spark is an example of such big data platform, however it is not restricted to only this combination.



3 Results

The models are simple and fast in execution. As false positives have high cost of investigation by human, entire methodology was tuned and optimized to provide optimum false positives. Below graph shows individual rate of false positives of origin and destination number detection. However when combined, combined mechanism provided <5% false positive.

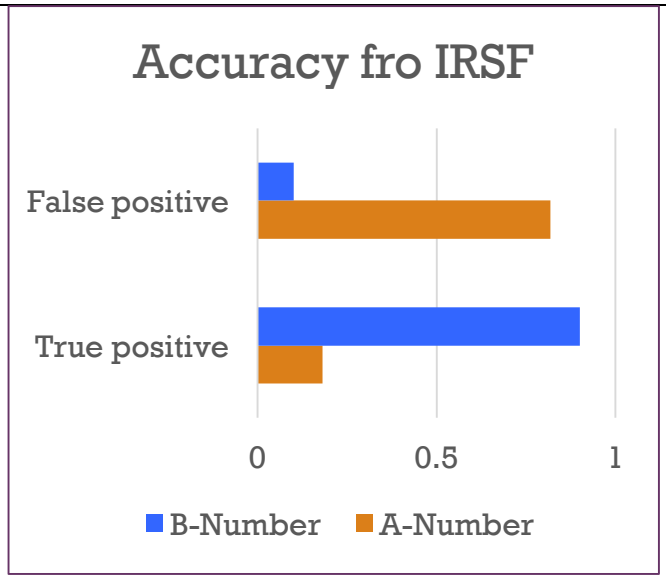


Figure 3-1True Positive versus False Positive for A(Origin)-number and B(Destination)-number

4 Learnings

#Simplicity is the key – Simple models are efficient and explainable. When multiple such models are combined, great results can be achieved.

#Exploratory Data Analysis – This is the step which does all work. Proper visualization gave us insight to deal with the problem and come up with simple engineered features.

#Break problem into small chunks – Once that's done, it's much easier to solve. For example, labeled data was not given to us. So we found Anomaly Detection technique to label the anomalies and then use those patterns to build supervised models. This made our models even better, as we could detect novelty patterns.

5 References

[1] Communications Fraud Control Association – "2013 Global Fraud Loss Survey", accessed from "http://www.cfca.org/fraudlosssurvey/"

[2] Constantinos S. Hilas and John N. Sahalos – "User Profiling for Fraud Detection in Telecommunication Networks", accessed from " http://icta05.teithe.gr/papers/69.pdf"

[3] Varun Chandola et al – "Anomaly Detection : A Survey" , accessed from "http://www.dtc.umn.edu/publications/reports/2008_16.pdf"