Machine Learning and the Evolution of Employee Fraud Detection

David Speights, PhD
Chief Data Scientist, Appriss

May 2017
Agenda

• Current state of exception reporting
• Collective intelligence – combining the knowledge gained from many retailers
• What is machine learning, predictive modeling, artificial intelligence, and deep learning?
• Parallels to credit card fraud detection
**Current Flow for EBR Investigations for a Single Query**

- **EBR Query on POS Data**
  - Investigation/Review of Exceptions
    - **Success**
      - Found Something to Correct/Termination
    - **Failure**
      - Found Nothing to Correct
  - Machine learning exploits the differences between successes and failures

- **Queries tend to focus on:**
  - Transaction anomalies, unusual sequences of transactions, etc.
  - Usually based on a small number of variables

- **Machine learning will use hundreds of variables to identify patterns of fraud based on experience from many different retailers and has the ability to learn/evolve over time as new fraud scams arise**
Loss Prevention Teams will have a Collection of Queries

Looking for Fraudulent or Problematic Employee Activity

How this feeds into machine learning

- Queries are converted into a collection of variables
- Query based variables are combined with many other variables:
  - Other retailers successful queries
  - SRAs (sales reducing activities)
  - Control variables: tenure, position, store location
Collective Intelligence
The Best Machine Learning Models will Use Many Retailer’s Strategies

Combining retailer strategies leads to better detection
- Variables which result from shared questions
- Models built on many successful strategies
Appriss Retail’s Base for Collective Intelligence
Sysrepublic + The Retail Equation Serves 90+ Retailers
Developing a Machine Learning System

Data Breadth

• Number of employees in database = 5+ Million
• Number of transactions = 45+ Billion
• Number of stores = 100,000+
• Incarceration data and the relationship with Appriss Public Safety = 125+ Million Incarcerations (US)

Model Development Efforts for Machine Learning Model Development

• Currently soliciting retailers to be involved in testing/model refinement
• 3 major U.S. retailers now participating
• Number of terminated employees used in modeling 10,000
• 3,000+ variables available for modeling
• Slated for general availability in late 2017
Artificial Intelligence, Machine Learning, and Predictive Models...

**Artificial Intelligence (AI)**
- Computer software/algorithms that emulate human intelligence or make automated decisions
- Predictive models and machine learning are tools/subsets of AI

**Predictive Models**
- Using data to estimate models which make predictions
- Originated by statisticians

**Machine Learning**
- Using data to “train” models which make predictions
- Originated by computer scientists

**Deep learning**
- Uses highly complex model structures to learn patterns in the data without any human intervention.
- Very hard to fit these models.

**Dynamic/Recursive Learning**
- Systems that re-fit or update the machine learning/predictive models dynamically as new data presents itself

Machine learning and Predictive models are very close to being the same thing
From Queries to Machine Learning
Variable Derivation is the Foundation

Objective – create a comprehensive list of model variables which can best predict problematic employees

Survey of Many Retailers top questions

Employee Fraud Study with Retailer #1

Employee Fraud Study with Retailer #2

Created list of core variables for model development

Key Variable Categories
- Voids
- Discounts
- Employee transactions
- Refunds
- Collusion
- No Sales
- Key sequences of transactions

More than 3,000 variables created
Machine Learning Approach
Information Flowchart

Employee Variables
- Employee Transactional Histories
- SRAs (Voids, Returns, Markdowns)
- Consumer / Employee Collusion Patterns
- Employee Self Run Transactions
- High Risk Behavioral Patterns

Decision Engine
- Machine Learning Models

Employees processed through predictive models continuously to evaluate risk

Employee Fraud Risk Score (000 to 999)
Employee Deviance Detection

A machine learning based approach will have a higher success rate than traditional approaches.
Dynamic Learning

Employee-based variables
3,000+

Known successful investigations
across many retailers

Machine Learning/
Predictive Models

Cases to Investigate

Feedback loop from investigations improves models & keeps up with new fraud

Feedback loop from investigations improves models & keeps up with new fraud
Key Components of Newer Technology
Identifying Variables’ Relationship to Target

Raw Data → Derived Variables → Mathematical Equation

Prob(Target)
No Fraud/Fraud

Traditional Methods
- Highly Manual Process
  Variable transformations and interactions a manual iterative process
- Simpler Equations

Machine Learning
- Still a manual process, but interactions and transformations handled automatically
- More Complex Equations

Adding Deep Learning
- Similar structures to machine learning, but attempting to work from raw data directly (idea is for no manual processing)
- Extremely Complex Equations

Requires more computing power and far more example targets
Example Problem Below:
• Non-linear relationship with interaction.
• Traditional methods require analyst to discover these facts

Machine Learning Automatically fits this relationship:
• Machine learning models detect complex relationships automatically
Employee Deviance: Predictive Modeling Evolution (credit card industry versus retail)

credit card fraud detection

Rules/Queries

Predictive Models/Machine Learning
- Neural Networks
- Extreme Gradient Boosting
- Regression Modeling
- Hidden Markov Models
- Random Forests

employee fraud/deviance detection

Rules/Queries

Predictive Models/Machine Learning
Machine Learning and the Evolution of Employee Fraud Detection

David Speights, PhD
Chief Data Scientist, Appriss

May 2017