Exploration and exploitation in deciding what to audit

Roman Chychyla

Alexander Kogan

Rutgers, The State University of New Jersey

2013

Problem Description

- **Problem**: Identify irregular transactions in a multi-period setting.
- **Challenges**:
  - Lots of transactions.
  - There is a cost to investigating a transaction.
  - Limited audit resources.
- **Solution**:
  - **Traditional:** choose a random sample from the population of all transactions.
  - **Modern:** use analytical (learning) models to identify suspicious transactions from the population of all transactions.

Statistical Model's Challenges

- **One-sided feedback** – model only learns from the previous transactions that were identified by it as suspicious and were investigated.

Transaction	True Nature
x-x-x	Irregular
y-y-y	Non-irregular

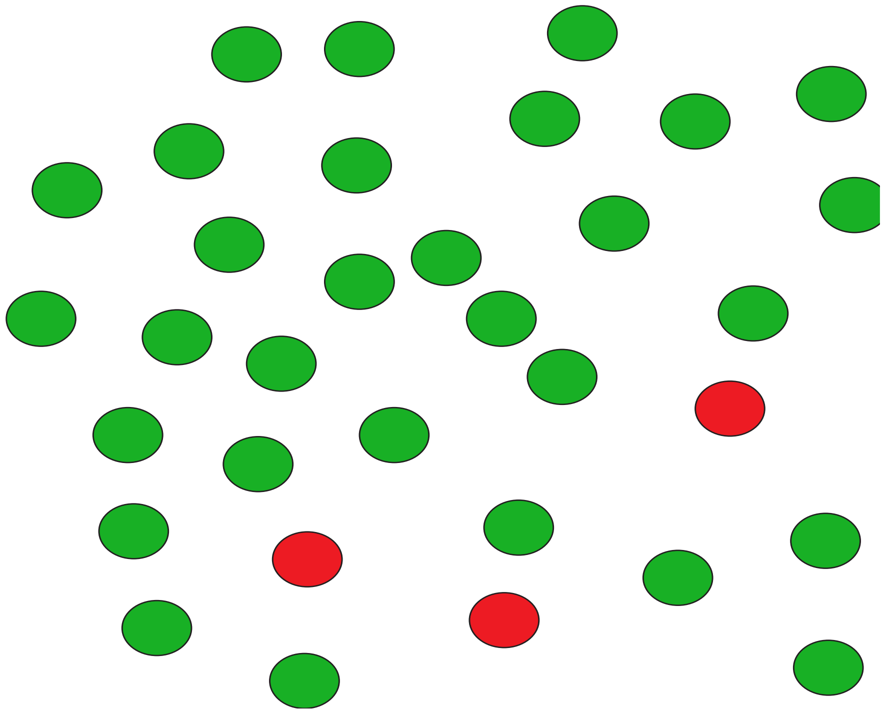
Previously investigated transactions

Transaction	True Nature	Investigate?
x-x-y	Irregular	Yes
z-z-z	Irregular	No
y-y-z	Non-irregular	No
z-z-y	Irregular	No

New unobserved transactions

Statistical Model's Challenges

- **Unbalanced data set problem** – number of irregular transactions is relatively small



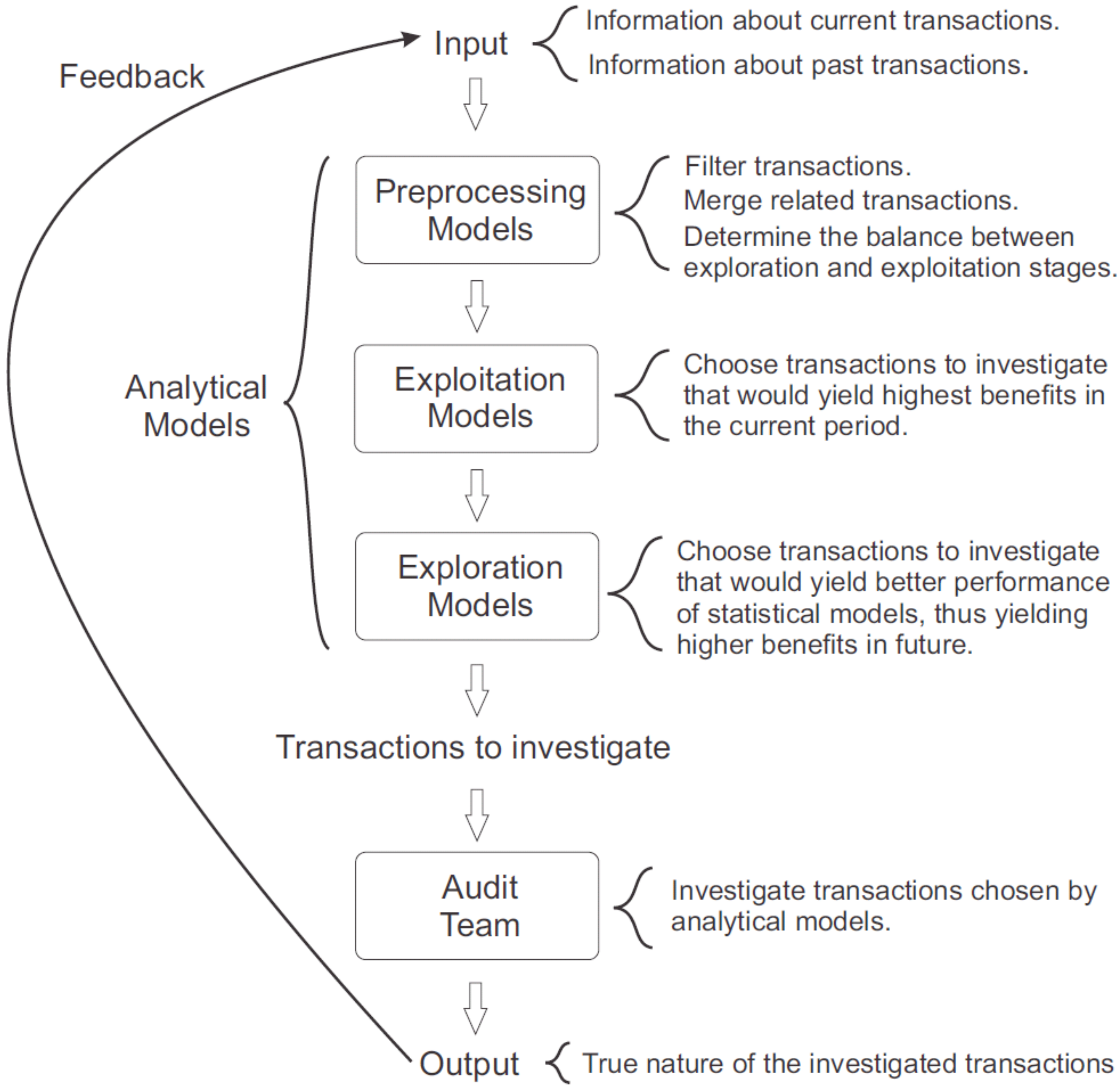


Figure 2: The exploration and exploitation framework for improving analytical models.

Empirical testing

Data sets:

1. **Multinational bank credit card data**
   - private data set
   - 500,000 observations
   - 5,000 (1%) irregular observations
   - 11 variables
   - observation is considered to be irregular, if the credit card was canceled by the bank
   - credit limit is assumed to be the value of a loss, if the observation is irregular
2. **U.S. census data**
   - public data set (used in 1999' KDD Cup competition)
   - 200,000 observations
   - 7,881 (3.94%) irregular observations
   - 13 variables
   - observation is considered to be irregular, if the person has a graduate degree (Master or PhD)
   - age of the person is assumed to be the value of the potential loss

Empirical testing

- Statistical models:
  1. Logistic regression
  2. Support Vector Machines (SVM) with linear kernel (LIBSVM implementation)
- Number of transactions in a period: 1000
- Audit capacity: 100 (10%)

Credit Card Data Results

		Exploration/exploitation models			
	Normal model	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 1$
Logistic Regression					
MRPL	11.56%	21.35%	23.37%	24.58%	23.90%
Difference	0%	9.79%	11.81%	13.02%	12.34%
Linear SVM					
MRPL	15.89%	17.20%	16.62%	16.26%	16.24%
Difference	0%	1.31%	0.73%	0.37%	0.35%

Table 2: Credit card data testing results as measured by the Mean Relative Prevented Loss (MRPL) in percentage. The difference row indicates the difference in MRLP between the exploration/exploitation models and the normal model. Higher values are better.

Census Data Results

		Exploration/exploitation models			
	Normal model	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 1$
Logistic Regression					
MRPL	22.53%	35.44%	35.93%	34.67%	33.24%
Difference	0%	12.91%	13.4%	12.14%	10.71%
Linear SVM					
MRPL	20.92%	28.68%	28.76%	25.44%	22.47%
Difference	0	7.76%	7.84%	4.52%	1.55%

Table 4: Census data testing results as measured by the Mean Relative Prevented Loss (MRPL) in percentage. The difference row indicates the difference in MRLP between the exploration/exploitation models and the normal model. Higher values are better.

Thank You!