Procurement Card Fraud Detection Using Hidden Markov Models and Information Fusion: 
A Fusion Study

Abdullah Al-Awadhi and Deniz Appelbaum

Introduction

- Large Multinational Consumer Goods Manufacturer
- Manual Batch Fraud Detection System
- Required Customized Approach, Commercial CAATs insufficient
- Many vendors do not report items that were purchased
- CARLab created a supervised system as first phase - ILisa

 ILisa: Uses duplicate testing, key words, and hundreds of association rules
- Can drill down to desired level of testing
- Elicits the expertise of the auditors

Measure for Jan 2013 to April 2014

<table>
<thead>
<tr>
<th># of Transactions</th>
<th>741,710</th>
<th>Missing Purchase Item Information Data Set</th>
<th>199,528 (26% of total)</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Employee IDs</td>
<td>4532</td>
<td># of cards are 5600</td>
<td>4339 (95.74%)</td>
</tr>
<tr>
<td>Total $ in Original Currency</td>
<td>$157,115,194</td>
<td>$65,926,544 (42% of total)</td>
<td></td>
</tr>
<tr>
<td>Total # of vendors</td>
<td>101,900</td>
<td>Total of 41,238 (40.49%)</td>
<td></td>
</tr>
</tbody>
</table>

A Fusion Study

- ILisa can’t analyze the missing purchase item transactions
- New types of fraud need an unsupervised approach
- Using Hidden Markov Models embedded in a Belief Network with Dempster-Shafer:

 ILisa: Textual analytics (TA) on

Levels of Analysis:

1. Only Textual analytics (TA) on items/merchants
2. TA and MCC filtering
3. TA, MCC-filtering, and General Rules
4. TA, MCC-filtering, General rules and Rules specific to firm

Future Research

- Application of ILisa rules to the Information Fusion to detect false positives and negatives
- Missing Purchase Item Transactions to be analyzed with the Hidden Markov Model add-in
- Operational Efficiency of model needs to be tested
- Accuracy rate should be at 85%, based on previous research
- Missing Purchase Items is not atypical for the industry

The Data: 172 confirmed fraudulent transactions

The Data - ID # 3937 @ Petsmart
plant and maintenance employee!!

#1469, 6/21/2013, $21
63 total transactions for $4436

Rutgers Business School
Newark and New Brunswick
Securing Big Data Provenance for Auditors: The Big Data Provenance Black Box

By Deniz Appelbaum

Data Provenance
- Origin of Data
- Lineage of Data
- Log files
- Data life cycle
- Component of Data RELIABILITY

Big Data!!
- Velocity: streaming very quickly and incessantly
- Variety: textual, social media, financial, sensor, pictures, audio
- Volume: massive, petabytes of data
- External sources to the firm

Why is Data Provenance Critical?
- Detection Risk
  - Appropriateness (Quality of Evidence)
  - Sufficiency (Quantity of Evidence)
  - Relevance (What Does Evidence Tell the Auditor?)
  - Reliability (Can the Auditor Trust the Evidence?)

Data Provenance in Hadoop
- HadoopProv (Akoush et al, 2013)
- 10% temporal lag in Hadoop
- Provenance is not secure
- Black Box: provenance is write once, read only (Alles et al, 2004)
- Digital signatures via hashing in the Black Box provide ultimate form of security
- These digital signatures reveal if the provenance records have been altered - the ultimate Big Data Provenance Black Box!

The Big Data Provenance Black Box:
- Based on HadoopProv
- Applies Black Box for secure data provenance for auditors!

Figure 1: Provenance Capture in HadoopProv

Rutgers Business School
Newark and New Brunswick
The Implementation of Exploratory Data Analysis (EDA) on State Data

Desi Arisandi and Miklos Vasarhelyi

Introduction

GASB Concepts Statement No. 1:
Accountability is the cornerstone of all financial reporting in government. Accountability requires governments to answer to the citizenry—to justify the raising of public resources and the purposes for which they are used.

The volume of government financial data and the change of structure of financial report due to standards implementation are raising the possibility of information overload. The overwhelming and dynamic information can increase the difficulty to understand the underline information of financial statements.

Exploratory data analysis (EDA) is one of contemporary methods that can assist to mine the information within substantial amount of data. The analysis can reveal every changes and dramatically flow of financial resources that are disclosed in the financial statements.

Methodology

Data
- Financial data: CAFR of 51 of states level entities in United States and cover 10 years periods (2004-2013)
- Non-Financial data:
  - Crime statistic (Source: FBI)
  - Leadership change (Source: Entity’s website)
  - Transparency Award (Source: GFOA)
  - Public employee corruption Conviction (Source: DOJ)

Analysis
- Ratio Analysis
- Cluster Analysis
  - Initial cluster based on the classification of National Statistic Bureau (Geographical based)
  - K-Mean cluster method

Expected Result
- Overall trends
- Cluster result based on the financial and non-financial information
- Anomaly Analysis

Results and Analysis

Cluster Analysis:
- Washington DC
  - Washington DC is not a state and based on U.S. Constitution this district is based on the Congress jurisdiction hence the deviation of law and budgetary structure with other states.
- Arizona
  - Low percentage of Government Service Revenues/Total Program Revenues: AZ average = 0.05, whole state average = 0.15
  - The state of Arizona is sinking in debt (State Data Lab, 2014). The State financial report shows $4.2 billion shortfall represents compensation and other costs

Conclusion
Trends and potential anomalies can be detected by implementing EDA. Furthermore, the graphical-based result can also support users understanding of the information.

Government Financial Reports

The Comprehensive Annual Financial Report (CAFR) is a thorough and detailed presentation of the state’s financial condition. It reports on the state’s activities and balances for each fiscal year (GASB, 2014).

CAFR is presented in three sections:
- Introductory section
- Financial section:
  - Required Supplementary Information (RSI)
  - Basic financial statement
  - Notes to financial statement
  - Audit Report
- Statistical section

In general the government activities can be classified into governmental, business, and fiduciary

From Alaska MD&S (2004-2006):
- Change of leadership
- The state total debt increased due to the development of sport fishing infrastructure and international airport
Legal Risk Prediction Model for Credit Card

Feiqi Huang
Qi Liu
Miklos Vasarhelyi

Introduction

Legal risk is special and important for banking and finance. Companies are usually stuck by lawsuit which may cause extremely large expense. Meanwhile, customer’s lawsuits against bank is a serious problem. Reports show larger global banks’ legal tab is more than $100 billion. In addition, unlike most other operational risks, legal risk cannot be traded away in any market. However, legal risk is not like other operational risks which have been fully analyzed by quantitative analysis.

Prior literature claims that legal risk is an indicator of the weakness of internal control and reflection of bad operational performance in the future. SAS No.109 requires auditors have a sufficient understanding of the entity, environment and internal. Besides traditional audit which is backwards or retroactive, a new audit focus is forward looking or predictive. In business area, predictive models is a common way to exploit patterns found in historical data to identify risks and opportunities.

To the best of our knowledge, there is no existing literature that focuses on legal risk prediction.

Methods & Measurements

In the process of building prediction models, authors use SAS to preprocess data and employ SPSS Modeler to build prediction models.

In the learning process, nine supervised algorithms are used to build prediction model: C5.0, CHAID, decision List, C&R tree, QUEST (Quick, Unbiased, Efficient Statistical Tree), Bayesian Network, Discriminant, Neural network and Logistic regression.

Data reflects less than 2% of clients have ever sued the bank. This feature (imbalance data) leads the predictive accuracy, the common measure of performance of prediction model, might not be appropriate. Receiver Operating Characteristic (ROC) curve, Recall and Precision are measurements of models performance.

Data Description

The data sets is related credit card business from a Major South American financial group. Cardholder information data describes each account holders’ personal information which contains 289 variables and 67,049,047 observations. Lawsuit data records each lawsuit case’s information and contains 256 variables and 1,495,673 instances. Complain data shows clients’ complains records, which has 26 variables and 1,116,386 records. Default data contains 50 variables and 53,224,215 observations and presents credit card holders’ default information. The last dataset is about Credit card restriction. It has 27 fields and 197,950,335 records. The combined data set contains 42,235,966 distinct clients and 598,431 of them (1.4%) have sued the bank.

Prediction Model

Trained by balanced training data, the best four algorithms are C5.0, Neural network, CHAID and logistic regression, which achieve 99.1%, 97.5%, 97.4% and 94.9% area under ROC curve respectively. When we applied the best model on testing data set, the C5.0 model achieves 95.63% Recall rate and 18.91% Precision rate. According to this model, 26 variables are used in the decision tree. The depth of the tree is 24 and contains hundreds of rules. The most five important variables in C5.0 model: number of inactive cards, indicator about whether the client’s cards are blocked, number of active cards, age and indicator about whether the credit card is restricted.

Future Work

- Minimizing cost by cost matrix
- Dimension reduction
- Prediction potential conspired lawsuits
- Analyzing causes of lawsuits
The Vision

Audit App Selection

- Audit apps are formalized audit procedures performed through computer scripts.

Example – Caseware Marketplace

A Potential Problem

- The increase in number and variety of audit apps complicates the app selection process.
- Auditors, especially those with less experience, will likely desire guidance or assistance when selecting apps for specific audit clients.

Standard-based Filtering

- The system filters audit apps by industry, business cycle, account, assertions, and audit objectives

Recommended based on audit standards

<table>
<thead>
<tr>
<th>Industry</th>
<th>Finance</th>
<th>Manufacturing</th>
<th>Education</th>
<th>IT</th>
<th>Utilities</th>
<th>Services</th>
<th>B2B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Procurement</td>
<td>Sales</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sales Revenue</td>
<td>Accounts</td>
<td>Cash</td>
<td>Sales Returns and Allowances</td>
<td>Accounts Receivable</td>
<td>Completeness</td>
<td>Occurrence</td>
<td>Classification</td>
</tr>
</tbody>
</table>

Auditors & Clients’ Effects

Recommendations based on auditors’ preferences

<table>
<thead>
<tr>
<th>Auditor</th>
<th>Audit App 1</th>
<th>Audit App 2</th>
<th>Audit App 3</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auditor 1</td>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Auditor 2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Auditor 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Recommendations based on audit clients

Methodology

- The immense number and variety of tests necessitate a system that can assist auditors in discovering the most appropriate audit apps.

Final Recommendation

- Generate a score for each audit app by combining the predicted rating based on the auditor preference and the predicted rating based on audit clients
- Score = \( \delta \times \text{predicted rating from auditor} + (1-\delta) \times \text{predicted suitability of the audit app} \)
- Audit apps with high scores will be recommended to the auditor in a particular audit engagement

Conclusion

- In this paper, we propose an audit app recommender system to provide digital suggestion for auditor.
- By analyzing audit environment and auditors’ historical behaviors, the recommender system can provide “personalized suggestions” for a particular auditor in a particular audit engagement.
A Novel Method for Outlier Detection

Paul Byrnes

Abstract
Organizational fraud is a growing problem for which solutions are needed. In fact, both companies and auditors are becoming more active in addressing this problem. In alignment with this, outlier detection can assist with the fraud discovery process.

In this research, a unique, automated multivariate outlier detection method is developed and implemented. The approach relies upon four recognized measures that are used in both an individual and aggregated manner to identify anomalous objects. Individually, each measure separately determines the extent to which an object deviates from a representative data point (i.e. median). In the aggregate, all measures are combined to produce an overall outlier score for each record. Preliminary results suggest that the outlier scoring method is useful for assisting with outlier detection in numerically represented data.

Introduction
Outliers have historically been described in a variety of ways. For example, Hawkins (1980) referred to an outlier as “an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism”. Barnett and Lewis (1994) described an outlier as “an observation which appears to be inconsistent with the remainder of that set of data”.

Irrespective of specific definition, an outlier, exception, or anomaly can be perceived as an object that is substantially different from other objects in the set to which it belongs. Outlier detection is a method for capturing those objects that are notably different from others (Zimek et al., 2014).

Background
In this study, outlier detection entails preliminary considerations. First, a suitable measure of central tendency is needed. While the mean might seem an obvious choice, it is only appropriate when the data approximates a normal distribution. However, data often deviates from this structure. Fortunately, the median is applicable in any case.

Second, the metric set to be used in anomaly detection is an important consideration because it heavily influences outcomes (Chandola et al., 2009). Zimek et al. (2014) caution that two measures of a particular type will tend to be more highly correlated than two metrics of differing types. They propose the use of ensembles in outlier detection whereby more than one measure is deployed. Given this, multiple indicators are selected for this study.

An array of potential distance measures are available for consideration, including Manhattan, Minkowski, Euclidean, and Mahalanobis. Furthermore, similarity measures exist such as the Jaccard Coefficient, Cosine Similarity, and the Tanimoto Coefficient (Tan et al., 2005). After evaluation of strengths and weaknesses, four measures are chosen: 1) Mahalanobis distance, 2) Euclidean distance, 3) Cosine similarity, and 4) Tanimoto coefficient. The last two are converted to dissimilarity measures. In this way, larger measurements always indicate higher outlier likelihood.

Method
In addition to initially examining each metric individually, a mechanism is used to aggregate outcomes for all four measures, thus producing an outlier score for each object. In achieving this, the results for each measure are normalized on a (0, 1) scale, which means that the maximum value for a particular metric is 1.

The outlier score for each object is computed as the sum of its normalized values for all measures. Consequently, the outlier score for a given record must lie between 0 and 4 (i.e. (0,4))

The object having the highest outlier score is deemed as most suspicious. Conversely, the record with the lowest outlier score is viewed as most similar to the median value, and, therefore, least problematic. To facilitate efficiency, the outlier detection process is substantially automated and produces rudimentary visualizations as well as an output file that can be readily explored in more sophisticated visualization software packages. In the following section, Tableau is used for image generation.

Analysis/Results
To gain initial insight, a series of plots representing all pair wise combinations of measures is created.

In each image, the median is at the origin, and objects farther from this are more anomalous. For example, the circled object in the upper right graph is identified as second most different from the median in terms of both Mahalanobis Distance and Cosine dissimilarity. This same object is again circled in the middle lower plot. While it remains far from the origin, its status as an outlier is less obvious in terms of Euclidean distance and Tanimoto dissimilarity. Next, outlier score visualizations are created to offer more specific insights.

In the above dashboard, records with the most significant outlier scores are emphasized. For instance, 1106838 has the highest outlier score (i.e. 3.511), indicating it is most different from the median. In fact, all records with outlier scores above 3 are particularly suspicious (see box plot view). An initial data review process indicates that, as outlier scores increase, records become increasingly different from the median result.

Conclusion
Anomaly detection is becoming more important. In this study, a novel outlier detection method is developed and implemented. While this study is still evolving, initial results show that it can successfully identify and prioritize outlier candidates in numerically represented data.
Internal Audit Scheduling Project
Factors affecting internal audit time duration and audit planning optimization
Qiao Li, Junming Liu, Miklos A. Vasarhelyi

Introduction
Interim Objectives:
- Which factors affect the elapsed time differences in completing audit engagements
- Budget Management
- A new risk-based audit planning/scheduling model

Preliminary Hypotheses
- Hypothesis 1: does audit elapsed time vary significantly with audit entity category?
- Hypothesis 2: do audits with higher risk levels need more audit time?
- Hypothesis 3: does the quarter affect audit time?
- Hypothesis 4: does the number and types of risks affect audit time?
- Hypothesis 5: do the issues reported after control affect audit time?

Problem Structure

Factors correlated with audit elapsed time

Before optimizing audit scheduling process, the reason why audit engagements take long time to issue is considered. In order to figure out which factors have significant effects on audit duration, some of the dimensions are observed and explored first: quarter the audit starts, audit rating, the level 1 entity or business line, the number of Critical, High, Medium, or Low risks, the number of reported issues, size of the audit (total number of budgeted hours), number of staff, and titles of staff working on the audit, etc.

Since the actual audit duration is unknown in the data set, it is defined as = last booking date of all staff involved in the engagement - first booking date of all staff involved.

The following graph shows the percentage of audits that were rated as satisfactory is declining across years.

Correlation
Number of issues reported is more correlated with audit duration, and the correlation is higher if considering audit hours than days.

Factors Analysis
Global risk score & audit time duration (days)
- the assumption that higher risk audits need more time applies to some business lines, but not to all.

Quarter of which the audit starts & audit time duration (days)
- shape of the four distributions are different, but engagements that start at quarter 4 actually were not completed in shorter period even the time is much closer to the end of a year.

Allocation of staff job position & business line

Factors Analysis
Quarter to start & delay ratio
— delay ratio = delay of days start working after planning completed/total days; for some business lines, delay ratio is higher if the audit engagement starts in early quarter (Jan. or Apr.)

Continuing Work
- Finding out reasons of extreme and unusual
- Weighting risk assessment scores for audits
- Using classification and regression methods to predict whether an engagement will go beyond time limit
The Original “Red Book” and Its Expanded Conceptualization

<table>
<thead>
<tr>
<th>Vasarhelyi &amp; Halper (1991)</th>
<th>Expanded conceptualization</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPAS: Foundations effect</td>
<td>Several, corporate, experiential, efficient</td>
<td></td>
</tr>
<tr>
<td>Measuring Metrics</td>
<td>Stands of measurement scales, standards of measurement</td>
<td></td>
</tr>
<tr>
<td>Creating a model</td>
<td>Standards of measurement scales, standards of measurement</td>
<td></td>
</tr>
<tr>
<td>Relating Analytics</td>
<td>Representational equation, continuous equation</td>
<td></td>
</tr>
<tr>
<td>råter</td>
<td>Continuous equation, continuous equation</td>
<td></td>
</tr>
<tr>
<td>Dynamics</td>
<td>Self-configuring, self-organizing</td>
<td></td>
</tr>
<tr>
<td>Clustering</td>
<td>Self-configuring, self-organizing</td>
<td></td>
</tr>
<tr>
<td>Transaction level</td>
<td>For automatic broad detection and transaction-level correction</td>
<td></td>
</tr>
<tr>
<td>Atoms (of events)</td>
<td>Measurement (partial data acquisition)</td>
<td></td>
</tr>
<tr>
<td>Data</td>
<td>Introducing external, continuous, connected, benchmarking</td>
<td></td>
</tr>
<tr>
<td>Dimensions</td>
<td>Probabilistic data, stochastic, probabilistic data, stochastic data</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>Continuous Control, Monitoring, CCM</td>
<td></td>
</tr>
<tr>
<td>Risk</td>
<td>Risk: (CRM &amp; C)</td>
<td></td>
</tr>
<tr>
<td>Compliance</td>
<td>Compliance: (C)</td>
<td></td>
</tr>
</tbody>
</table>

Evolving Database Structures and Their Audit (Expanded from Vasarhelyi & Halper 1991)

- Database: Documentation
- User query: Auditor query
- Transaction: Examine levels, capture sample transactions
- Duplex: Sorting and listing, logical analysis and indexes
- Field analysis: Paper oriented, software based
- Security issues: Physical, Access hierarchies
- Restat & Recovery: Plan analysis, Direct access
- Database interfaces: Reorganization, reconciliation and transactional integration

CA Redefined

A continuous audit is a methodology that enables independent auditors to provide assurance on a subject matter, for which an entity’s management is responsible, using a continuous opinion schema issued virtually simultaneously with, or a short period of time after, the occurrence of events underlying the subject matter. The continuous audit may entail predictive modules and may supplement organizational controls.

The continuous audit environment will progressively automate with auditors taking progressively higher and more progressive judgment functions. The audit will be by analytical, by exception, adaptive, and cover financial and non-financial functions.

The New CA

The major changes to CA that are emerging and should be permeating the audit environment and hopefully standards are:
1. Progressive adoption of a standard data interface to allow for the usage of assertion and analytic based “apps.”
2. The need to incorporate Exploratory Data Analysis into extant audit methodology. Liu (2014) proposes such a step where she expects intelligent modules to be able to interact with a wide variety of data sources.
3. Progressive impounding of audit apps into the operating environment.
4. The evolution of an audit ecosystem with progressive level of automation over financial and non-financial systems.
5. An environment rich of software agents (krons and daemons) activated by conditions or timing and acting both over data received (inputs) from upstream system, data entry, and automatic capture and examining data to be fed to downstream systems in a predictive audit mode.

Evolving Database Structures and Their Audit (Expanded from Vasarhelyi & Halper 1991)

- Database: Documentation
- User query: Auditor query
- Transaction: Examine levels, capture sample transactions
- Duplex: Sorting and listing, logical analysis and indexes
- Field analysis: Paper oriented, software based
- Security issues: Physical, Access hierarchies
- Restat & Recovery: Plan analysis, Direct access
- Database interfaces: Reorganization, reconciliation and transactional integration

Envisaged within An Audit Ecosystem (Vasarhelyi and Kozlovski 2014)

<table>
<thead>
<tr>
<th>Data Assuranc</th>
<th>Controls</th>
<th>Compliance</th>
<th>Risk Monitoring and Assessment</th>
<th>Operations and Monitoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who uses</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Management</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Audit (internal or external)</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Investors</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regulators</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purpose</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnostic</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Predictive</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Proactive</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Primarily performed by</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Automation</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Manual</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

Incorporating EDA (Liu, 2014)
The application of an Audit Ecosystem concept to an enterprise system

Stephen Kozlowski

Introduction
An audit ecosystem provides a technology-driven, self-sustaining audit function for firms and organizations of varying sizes and configurations. The ecosystem will leverage the digital capabilities in place at the firm. Many firms have implemented computer-based accounting systems ranging in size from PC-based packages to tailored ERP systems.

The advent of the internet provides a platform that allows for the collection of varied types and large amounts of data. Although not necessarily sourced from within the firm itself, certain forms of this data may provide insights to the firm’s operations that can complement the audit function.

This current research investigates the applicability of an ecosystem approach in conjunction with an ERP system.

Continuous Audit Model
Data sources:
- Client ERP system consisting of transactional data and logs
- Other client automated systems
- Client’s manual systems
- External data that will require design of an appropriate data receptacle

Financial data will be standardized to comply with Audit Data Standards

Analytic tools will analyze system logs to identify data paths to develop meta evidence to address audit risk

A tailored audit plan will be developed that considers:
- Industry
- Audit experience
- Auditor characteristics
- Incorporates analytic results of system logs and exogenous data

An evidence suggestion system will be developed to model the value of audit evidence

Appropriate audit applications will be identified and launched

Applications will be launched to interpret the results

Audit findings will be presented and further actions will be indicated

Proposed Design

Related Activities
- ERP data for A/R, A/P, and G/L was obtained from one NFP client, typical CA/CM techniques were applied, and the results were provided to the client
- Payroll and H/R data was provided from a second client, appropriate audit tests as requested by the client were applied, and results provided to the client using spreadsheets and dashboards
- An automated testing routine was developed and implementation is underway by the client
- Project planning is underway with a third client who has indicated they will provide Payroll, G/L, and A/P data for analytical purposes
Auditing Analytical Procedure Techniques: Does Process Mining Complement or Substitute Data Mining?

Tiffany Chiu and Miklos Vasarhelyi

Introduction

Unlike traditional auditing analytical procedure, process mining of event logs provides a new aspect for audit in the way that this technique analyzes and processes transaction data for each and every business event instead of relying on only a sample of the population. Prior literature indicated that both process mining and data mining techniques can add value and improve the performance of analytical procedures in auditing. However, it is still not clear whether process mining of event logs and data mining techniques should be applied together in a complementary fashion or process mining of event logs could replace data mining techniques.

This study aims at analyzing and comparing the performance of process mining and data mining techniques in auditing analytical procedure using Volvo IT dataset, and distinguish whether process mining complements or substitutes data mining technique.

Process Mining of Event Logs

Process mining refers to the usage of event logs to analyze business processes. There are four characteristics that must be extracted from each event in the system in order to analyze the data:

<table>
<thead>
<tr>
<th>Characteristics of Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Activity</td>
</tr>
<tr>
<td>(2) Process Instance</td>
</tr>
<tr>
<td>(3) Originator</td>
</tr>
<tr>
<td>(4) Timestamp</td>
</tr>
</tbody>
</table>

Prior studies proposed that when utilizing process mining techniques to analyze the information from event logs, five different types of analysis can be performed in process mining:

- Process discovery
- Conformance check
- Performance analysis
- Social network analysis
- Decision mining and verification

Literature Review

- Application of Process Mining in Audit
  - Jans et al. (2009) proposed a framework for reducing internal fraud risk based on process mining event logs.
  - Jans et al. (2013) discovered that process mining can add value to audit as it enhances the effectiveness of fraud prevention, especially when auditees are made aware of event logs.
  - Jans et al. (2014) applied process mining of event logs in auditing analytical procedures, and successfully detected anomalous transactions that traditional auditing analytical procedures may fail to discover.

- Application of Cluster Analysis in Audit
  - Thiprungsri (2010) applied cluster analysis to group transactions of transitory accounts; results indicated that cluster analysis is useful for detecting anomalous transactions in audit.
  - Thiprungsri and Vasarhelyi (2011) examined life insurance claims using clustering and proposed that cluster analysis is a promising technique that can be integrated into the concepts of continuous system monitoring and assurance.

Methodology and Dataset

This study applied process mining and data mining techniques, respectively, to analyze a real life Volvo IT dataset. The unsupervised learning algorithm (cluster analysis) – K-mean and Fuzzy Miner technique in process mining are employed to analyze and compare the data.

Volvo IT Problem Management

| Total Number of Process Instances (cases) | 819 |
| Total Number of Events                   | 2,351 |
| Problem Status                           | 3 |
| Problem Sub-Status                       | 5 |
| Problem Involved Action Owner            | 240 |

Future Research

Process mining may enhance the performance of “Audit by Exception” concept proposed by Vasarhelyi and Halper (1991). Audit by exception refers to the usage of CPAM in audit procedure so that the audit works will be focused on the alarm of exception gathered by the system on a continuous basis.

Application of Process Mining with “Audit by Exception”: An alarm will arise when purchase order is released without proper sign. The Figure below shows the example process; the flow chart is a procurement process extracted and revised from Jans et al. (2014).
Interactive Auditor Dashboard:
Application On Life Insurance
Basma Moharram and Miklos Vasarhelyi

What to Dashboard?
Our objective is to create an auditor dashboard to assist the auditor in designing and performing his audit plan. The first question we had to ask ourselves was what to dashboard. To answer this question we followed this approach: We start with a specific industry (Insurance). We break down into its main business cycles. We then break each business cycle into its main functions. For each main function we think of the possible assertions the auditor would want to test. In deciding the assertions we use the AICPA audit guide, audit plans, audit analytics, and audit apps.

Exploratory Analysis
The Chart shows the amount of Claim payments made to clients by company code. The auditor can filter for a specific range of payments. He can drill down from Company level, down to type of insurance, to type of product, until we go down to each single claim. Using this graph, an auditor will gain an understanding of the claim payments made by different companies.

Approvers Activities
A chart showing both the number of transactions authorized by a specific approver (The higher the number of claims, the bigger the size of square) and the total monetary value he authorized as claim payments (the higher the value, the darker the green). An auditor using this graph might be interested in the approver who approved the highest monetary value, or he might be interested in the approver who only authorized one single transaction (smallest square on the lower right corner). The auditor can right click any square to see the actual data.

Premium Outliers
Based on a RobustReg SAS model, the chart shows potential premium outliers in orange. The ones under the blue line is specially important as it shows that the company is collecting less premium than it should.
**APs and Disaggregated Data**

Kogan et al. (2010) compare the widest range of statistical models and find that VAR models and linear regression models tend to perform better than others. Additionally, previous literature indicates that disaggregated model (micro-level) is likely to deliver better performance than monthly, aggregated level models on segment or product line balance (macro-level) on APs. Knechel (1988, Dzeng 1994; Allen et al. 1999).

**H1:** Firm-wide sales expectations developed from disaggregated individual location model produce more accurate and more precise expectation than firm-wide sales expectation derived from aggregated firm-level models.

**H2:** Firm-wide sales expectations developed from daily disaggregated individual location model produce more accurate and more precise expectation than firm-wide sales expectation derived from weekly disaggregated individual location models.

**H3:** The model with both financial and non-financial information produces more accurate and more precise prediction than the model with only financial information.

---

**APs and NFI**

SAS No 56 (AICPA 1988) suggests Non-financial information (NFI) should be considered when performing APs, and also it can be used to evaluate risks and detect material misstatements (AICPA 2002, 2007). According to SAS 56 (AICPA 1988) during APs to develop expectations of accounts factors such as financial data from prior periods, client financial budgets, and industry information could be used. Especially, it recommends analyzing the relation between financial information and NFI.

**Method (1/2)**

1. Data

   The data employed in this research was obtained from one of the world-wide served audit firms. The targeted firm is a multiplications service firm with homogeneous operation in the world, but in this research only observations from the U.S. are used. A total 24 monthly observations are provided, and especially it is for about 2,000 operating unit locations from the fiscal year 2011 to fiscal year 2012.

2. NFI

   Weather information such as daily precipitation and maximum temperature is utilized as non-financial information because in particular retail industry sales amounts are likely to be affected by weather condition (Engle et al. 1986; Maunder 1973; Starr-McCluer 2000).

---

**Prediction Model**

<table>
<thead>
<tr>
<th>Level</th>
<th>Model Description</th>
<th>Model Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A: Models Without NFI</td>
<td>Weekly Vector Auto-regression</td>
<td>$\hat{y}<em>{t+1} = \beta_0 + \beta_1 y</em>{t} + \beta_2 y_{t-1} + \ldots + \beta_k y_{t-k}$</td>
</tr>
<tr>
<td></td>
<td>Daily Vector Auto-regression</td>
<td>$\hat{y}<em>{t+1} = \beta_0 + \beta_1 y</em>{t} + \beta_2 y_{t-1} + \ldots + \beta_k y_{t-k}$</td>
</tr>
<tr>
<td>Panel B: Models With NFI in a Firm-Wide Level</td>
<td>Weekly Multivariate Regression</td>
<td>$\hat{y}_{t+1} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$</td>
</tr>
<tr>
<td></td>
<td>Daily Multivariate Regression</td>
<td>$\hat{y}_{t+1} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$</td>
</tr>
<tr>
<td></td>
<td>Weekly Vector Auto-regression</td>
<td>$\hat{y}_{t+1} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$</td>
</tr>
<tr>
<td></td>
<td>Daily Vector Auto-regression</td>
<td>$\hat{y}_{t+1} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$</td>
</tr>
<tr>
<td>Panel C: Models in a Store level Data</td>
<td>Weekly Multivariate Regression</td>
<td>$\hat{y}_{t+1} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$</td>
</tr>
<tr>
<td></td>
<td>Daily Multivariate Regression</td>
<td>$\hat{y}_{t+1} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$</td>
</tr>
<tr>
<td></td>
<td>Weekly Vector Auto-regression</td>
<td>$\hat{y}_{t+1} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$</td>
</tr>
<tr>
<td></td>
<td>Daily Vector Auto-regression</td>
<td>$\hat{y}_{t+1} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$</td>
</tr>
<tr>
<td>Panel D: Models With NFI in Store level Data</td>
<td>Weekly Multivariate Regression</td>
<td>$\hat{y}_{t+1} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$</td>
</tr>
<tr>
<td></td>
<td>Daily Multivariate Regression</td>
<td>$\hat{y}_{t+1} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$</td>
</tr>
<tr>
<td></td>
<td>Weekly Vector Auto-regression</td>
<td>$\hat{y}_{t+1} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$</td>
</tr>
<tr>
<td></td>
<td>Daily Vector Auto-regression</td>
<td>$\hat{y}_{t+1} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$</td>
</tr>
</tbody>
</table>

---

**Method (2/2)**

3. Control Variables

   This study is extended by the studies of Kogan et al. (2010) and Allen et al. (1999). Basically, there are two kinds of models tested in this study- the multivariate regression models and the vector autoregressive models. The store level model is supposed to have about 2,000 predictors which are observations from the other stores on the models, but too many independent variables cause full rank issues. Therefore, only highly correlated predictors are selected by stepwise selection methods.

4. Evaluation of models

   $MAPE =$ Abs (actual value –predicted value)/ actual value

   Each model generates one-step ahead forecast by rolling forecast.

---

**Preliminary Results**

<table>
<thead>
<tr>
<th>Level</th>
<th>Model Description</th>
<th>Adj. R square</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A: Models Without NFI in Store level Data</td>
<td>Daily Multivariate Regression</td>
<td>0.6910</td>
<td>0.1548</td>
</tr>
<tr>
<td></td>
<td>Daily Vector Auto-regression</td>
<td>0.7490</td>
<td>0.1495</td>
</tr>
<tr>
<td>Panel B: Models With NFI in Store level Data (Single store test result)</td>
<td>Daily Vector Auto-regression</td>
<td>0.7512</td>
<td>0.1450</td>
</tr>
</tbody>
</table>

As far as empirical works with one of the stores located in Gastonia, North Carolina, the weather information plays important roles in explaining the sales account but doesn’t improve the accuracy of expectations significantly.
Log4Audit

Tatiana Gershberg and Miklos Vasarhelyi

**Traditional Audit Evidence**

AU Section 326 pertaining Audit Evidence specifies that auditors obtain audit evidence by “testing the accounting records”. The testing may include analysis, review, reproduction of “procedures followed in the financial reporting process, and reconciling related types and applications of the same information.” Accounting records do not suffice as audit evidence; thus, auditors seek other information to explain how this data was compounded. Knowingly, financial data reported for auditing purposes is consolidated data gathered from various ERP systems within the organization.

**Log4Audit**

Reaching “conclusions through valid reasoning” by auditors can be supplemented with artificial intelligence providing exact set of events that led to an accounting record being examined. Predictive and, furthermore, preventive audit (Kuenkarkaew and Vasarhelyi, 2013) implementation here is essential. The model also allows for customization of analytics: search engine and Index engine, for example, enable tuning of search for keywords in certain proximity, or adjusting verbosity or severity of logging.

**Model**

- ERP systems already utilize a logging framework.
- Audit Events Logger’s main function is to accept messages, accompanied by a date and time stamp, its verbosity and severity.
- Audit Indexing Engine serializes the data by indexing it. By tagging the data processed within Audit Indexing Engine (AIE), we speculate that AIE assists with structuring the data and producing higher quality analytics.
- Audit Big Data becomes a repository of indexed unstructured data that is accessed by a search engine in order to produce analytics that satisfy the needs of Business Intelligence tools and Real Time Monitoring to generate meaningful output that is further investigated by auditors.
- Implementing methodologies that lead to diagnostics, prioritization and evaluating of anomalies would streamline the auditing process cycles, leaving the only exceptional cases for human judgment.