Using Data Analytics to Detect Fraud

Other Data Analysis Techniques
Benford’s Law

- In a population of naturally occurring, multi-digit numbers, the first digit of the numbers will be distributed in a predictable way.

Data set conditions for application:
- Should describe the sizes of similar phenomena
- Should have no built-in minimum or maximum values
- Should have no pre-assigned (non-natural) numbers
- Should have more small items than large ones
## Benford’s Law

<table>
<thead>
<tr>
<th>Digit</th>
<th>1&lt;sup&gt;st&lt;/sup&gt;</th>
<th>2&lt;sup&gt;nd&lt;/sup&gt;</th>
<th>3&lt;sup&gt;rd&lt;/sup&gt;</th>
<th>4&lt;sup&gt;th&lt;/sup&gt;</th>
<th>5&lt;sup&gt;th&lt;/sup&gt; or Greater</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>11.97%</td>
<td>10.18%</td>
<td>10.02%</td>
<td>10.00%</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>30.103%</td>
<td>11.39%</td>
<td>10.14%</td>
<td>10.01%</td>
<td>10.00%</td>
</tr>
<tr>
<td>2</td>
<td>17.609%</td>
<td>10.88%</td>
<td>10.10%</td>
<td>10.01%</td>
<td>10.00%</td>
</tr>
<tr>
<td>3</td>
<td>12.494%</td>
<td>10.43%</td>
<td>10.06%</td>
<td>10.01%</td>
<td>10.00%</td>
</tr>
<tr>
<td>4</td>
<td>9.691%</td>
<td>10.03%</td>
<td>10.02%</td>
<td>10.00%</td>
<td>10.00%</td>
</tr>
<tr>
<td>5</td>
<td>7.918%</td>
<td>9.67%</td>
<td>9.98%</td>
<td>10.00%</td>
<td>10.00%</td>
</tr>
<tr>
<td>6</td>
<td>6.695%</td>
<td>9.34%</td>
<td>9.94%</td>
<td>9.99%</td>
<td>10.00%</td>
</tr>
<tr>
<td>7</td>
<td>5.799%</td>
<td>9.04%</td>
<td>9.90%</td>
<td>9.99%</td>
<td>10.00%</td>
</tr>
<tr>
<td>8</td>
<td>5.115%</td>
<td>8.76%</td>
<td>9.86%</td>
<td>9.99%</td>
<td>10.00%</td>
</tr>
<tr>
<td>9</td>
<td>4.576%</td>
<td>8.50%</td>
<td>9.83%</td>
<td>9.98%</td>
<td>10.00%</td>
</tr>
<tr>
<td></td>
<td>100.000%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>
Benford’s Law
## Application: Benford’s Law

<table>
<thead>
<tr>
<th>Excel</th>
<th>ACL</th>
<th>IDEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>▪ COUNTIFS()</td>
<td>▪ Benford’s Law command</td>
<td>▪ Benford’s Law command</td>
</tr>
</tbody>
</table>

### Tableau

- Calculated Fields
- LEFT, LOG, INT
Statistical Analysis

- Averages
- Standard deviations
- Highest and lowest values
- Absolute values
# Application: Statistics

<table>
<thead>
<tr>
<th>Excel</th>
<th>ACL</th>
<th>IDEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>▪ AVERAGEIFS()</td>
<td>▪ STATISTICS command</td>
<td>▪ Field Statistics</td>
</tr>
<tr>
<td>▪ MIN()</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▪ MAX()</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▪ ABS()</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▪ STDEV()</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tableau</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>▪ AVG, RUNNING_AVG, WINDOW_AVG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▪ SUM, RUNNING_SUM, WINDOW_SUM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▪ MAX, RUNNING_MAX, WINDOW_MAX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▪ MIN, RUNNING_MIN, WINDOW_MIN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▪ STDEV, WINDOW_STDEV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▪ STDEVP, WINDOW_STDEV</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Trend Analysis

- Compares data across time, location, operational unit, or other source of comparable information, and identifies outliers
- Identifies unexpected trends, or lack of trends where trends would be expected
Regression Analysis

- Statistical technique that uses a series of records to create a model relationship between a dependent variable and one or more independent variables:
  - Example: predicting the number of items manufactured based on amounts of materials and labor used
Regression Analysis

- Must have a pair-wise relationship between the two sets of data:
  - Each $x$ variable must have a corresponding and unique $y$ variable.
  - Determine which is $x$ (independent variable) and which is $y$ (dependent variable).
## Application: Regression

<table>
<thead>
<tr>
<th>Excel</th>
<th>Tableau</th>
</tr>
</thead>
<tbody>
<tr>
<td>▪ Analysis ToolPak</td>
<td>▪ Table calculations</td>
</tr>
</tbody>
</table>
Correlation Strength

- Correlation analysis allows a fraud examiner to evaluate the strength of the connection between two or more pieces of information.
- The strength of the relationship between two variables is indicated by the correlation coefficient ($r$):
  - $r$ ranges from -1 to 1.
# Application: Correlation

<table>
<thead>
<tr>
<th>Excel</th>
<th>IDEA</th>
<th>Tableau</th>
</tr>
</thead>
<tbody>
<tr>
<td>▪ CORREL()</td>
<td>▪ Correlation analysis (Statistics)</td>
<td>▪ CORR</td>
</tr>
<tr>
<td>▪ Analysis ToolPak</td>
<td></td>
<td>▪ WINDOW CORR</td>
</tr>
</tbody>
</table>

© 2018 Association of Certified Fraud Examiners, Inc.
Linear Regression

- Graphing two variables against each other provides a best-fit line:
  - \( y = mx + b \)
- Best-fit line can be used to predict \( y \) for any given \( x \).
Linear Regression Analysis Example

\[ y = mx + b \]

- \( y \) is dependent variable (total cost of gala)
- \( b \) is constant amount (fixed cost of gala)
- \( x \) is independent variable (number of guests)
- \( m \) is coefficient of \( x \) (marginal cost of each additional guest)
## Linear Regression Analysis Example

<table>
<thead>
<tr>
<th>Number of gala guests</th>
<th>Total gala cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>53</td>
<td>$21,672</td>
</tr>
<tr>
<td>81</td>
<td>$23,059</td>
</tr>
<tr>
<td>118</td>
<td>$32,848</td>
</tr>
<tr>
<td>157</td>
<td>$39,916</td>
</tr>
<tr>
<td>174</td>
<td>$40,317</td>
</tr>
<tr>
<td>202</td>
<td>$51,902</td>
</tr>
<tr>
<td>227</td>
<td>$63,522</td>
</tr>
<tr>
<td>269</td>
<td>$64,134</td>
</tr>
<tr>
<td>306</td>
<td>$82,083</td>
</tr>
</tbody>
</table>
Linear Regression Analysis Example
Linear Regression Analysis Example

\[ y = 237.78x + 4677.36 \]
Reasonableness Testing

- This testing involves using nonfinancial and financial data to predict a financial result.
- Example: annual membership dues x number of members = estimated annual membership revenue
- Accuracy depends on reliability of the nonfinancial data.
Reasonableness Testing

Kevin Forrester
Convicted Fraudster
Predictive Modeling

- This type of analysis involves extracting information from past data and using it to predict trends, behavior patterns, and future outcomes.

- Most predictive models generate a score, with a higher score indicating a higher likelihood of the given behavior or event occurring:
  - Example: credit scores
Qualitative Data Analysis

- Most data analysis techniques require the use of data in the form of numbers.
- Qualitative data analysis is the process used to collect and analyze unstructured data.
- It requires extreme organization, as data sets tend to be quite large.
- It is best used in conjunction with quantitative data analysis.
Textual Analytics

▪ Analysis of unstructured, text-based data for keywords that reveal patterns, sentiments, and relationships indicative of fraud
▪ Uses linguistic technologies, statistical techniques, and scoring algorithms to categorize and analyze textual data
▪ Requires:
  • Sophisticated software
  • Thorough understanding of the legal issues regarding employee rights and workplace searches
Textual Analytics Data Sources

- Emails
- Social media posts
- Instant messages
- Videos
- News feeds
Textual Analytics Data Sources

- Apps
- Voice files
- User documents
- Sales and marketing
- Presentations
Textual Analytics

- Emotive tone—derogatory, surprised, secretive, angry, worried, or confused communications
- Ethical behavior—harassing, secretive, cursing
- Concept, entity, relationship, or event extraction
- Text link analysis
- Social network analysis
Developing a List of Keywords

- Keywords depend on industry, relevant fraud risks, and data set.
- Use the Fraud Triangle as a reference.

Keywords depend on industry, relevant fraud risks, and data set. Use the Fraud Triangle as a reference.

Keywords:
- Deadline
- Quota
- Short
- Problem
- Concern

Reasonable
- Deserve
- Borrow
- Therefore

Override
- Write-off
- Adjust
- Discount

Financial
- Pressure
- Rationalization
- Opportunity
Visual Analytics

- Some data sets are so vast or complex that they are difficult to analyze using traditional means.
- Visual analytics can help identify patterns or anomalies in such data.
Tree Maps

<table>
<thead>
<tr>
<th>Accounting</th>
<th>Accounts Payable</th>
<th>Information Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Financial Misstatement</td>
<td>Fraudulent Disbursements</td>
<td>System Failures</td>
</tr>
<tr>
<td>Penalties and Fines</td>
<td>Thefts of Assets</td>
<td></td>
</tr>
<tr>
<td>Reputation Risk</td>
<td>Financial Misstatement</td>
<td>Natural Disasters</td>
</tr>
<tr>
<td></td>
<td>Fraudulent Disbursements</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thefts of Assets</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Research and Development</th>
<th>Sales</th>
<th>Administrative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reputation Risk</td>
<td>Fraudulent Disbursements</td>
<td>Theft of Assets</td>
</tr>
<tr>
<td>Penalties and Fines</td>
<td>System Failures</td>
<td></td>
</tr>
<tr>
<td>Natural Disasters</td>
<td>Theft of Assets</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reputatio Risk</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fraudulent Disbursements</td>
</tr>
</tbody>
</table>
Link Analysis
Geospatial Analysis
Timeline Analysis

- Highlight key times, dates, and facts.
- More readily determine a sequence of events.
- Analyze multiple or concurrent sequences of events.
- Track unaccounted-for time.
- Identify inconsistencies or impossibilities in data.
Conflict Checking

▪ Use of specialized software to identify potentially problematic relationships:
  ▪ Conflicts of interest on a company’s board of directors
  ▪ Relationships between potential new hires and business units
  ▪ Vendors that operate from addresses that previously have been associated with fraud
  ▪ Vendors that are not listed in any online commercial database or that are listed on government excluded party lists
Log-in Monitoring Software

- Used to identify anomalies in log-ins to company system:
  - Unusual times or days
  - Unusual locations
  - Excessive log-ins
  - Unauthorized log-ins
Employee Risk Ranking and Scoring

- Ranking and scoring bring together big data—both structured and unstructured data—to identify and monitor high-risk employees.
- All tests and analytics can be aggregated into a dashboard.
- A similar analysis can be performed for vendors or customers.
- Data privacy and employee rights are integral concerns.
# Employee Risk Ranking Example

1. **Keywords**
   - Percentage of keywords around pressure, opportunity, and rationalization in email and IM communications
   - Scaling: 3

2. **T&E analysis**
   - Ranking of T&E out-of-compliance hits and overall email scoring
   - Scaling: 3

3. **Sales activity**
   - Ranking of sales activity, field notes, and sales returns and allowances
   - Scaling: 4

4. **User activity**
   - Percentage of instances within period where custodian sends or receives information involving those outside of peer group, as identified through hierarchies
   - Scaling: 2

5. **Third-party risk**
   - Instances where employee is linked to high-risk third parties (e.g., customers, vendors, state-owned entities, etc.) in email, T&E, or sales activity
   - Scaling: 2

6. **Alias clustering**
   - Percentage of instances within that week where custodian sends or receives ESI involving at least one of their identified communicative aliases
   - Scaling: 3

7. **Emotive tone**
   - Percentage of instances where the employee sends or receives ESI with negative emotions (angry, frustrated, secretive, etc.) identified through linguistic analyses
   - Scaling: 5

<table>
<thead>
<tr>
<th>Custodian</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
<th>C7</th>
<th>Scaling C1</th>
<th>Scaling C2</th>
<th>Scaling C3</th>
<th>Scaling C4</th>
<th>Scaling C5</th>
<th>Scaling C6</th>
<th>Scaling C7</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, Week 1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>45</td>
</tr>
<tr>
<td>A, Week 2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>37</td>
</tr>
</tbody>
</table>
Transaction Risk Scoring

- Brings multiple analytics together
- Monitors and scores transactions based on number of breaches:
  - Round-dollar amounts
  - Incomplete invoice
  - Keyword in payment description field
- Can apply weights to certain tests to influence the overall risk score