FRAUD ANALYTICS AND DATA WAREHOUSING

Office of the Comptroller
State of Maryland

Andrew Schaufele; Director, Bureau of Revenue Estimates
Agenda

• Our Team
• Background
• Case with TBD Success
• Successful Case
• Lessons Learned
Team

- Deputy Comptroller
  - David Roose
- Bureau of Revenue Estimates
  - Project Management
- RAD – QRDT
  - Principal Users
- Teradata
  - Primary Contractor
- ASR Analytics
  - Primary Analytics Developer
Existing Processes:
QRDT
(Questionable Return Detection Team)
QRDT Processes – Pre Updates

• TY2013 Capacity: 130,000 returns
  • Stopped $23 million
• Myriad stand alone metrics
  • Ex: Refund/Withholding Ratio
• Aggregate Hit Rate: ~10% (90% False Positive)
  • Significant Hit Rate Variance Among Programs
TBD Case: Census Model (Geospatial Analysis)
Business Case & Proposed Solutions

- Fraud may be concentrated in specific geographic areas.
- Smallest geographic area available is ZIP code.
- Idea:
  - Geocode return to more finite areas.
  - Review returns from areas with anomalous return counts relative to history/census data.
Initial Solution

- Interactive Drill Down To Identify Fraudulent Refunds Report That Grouped Similar Anomalous Returns Together
- By Census Block
  - Zero Population Anomaly – Verify the block exists and is populated
  - Longitudinal Historical Filing Time – 95% confidence interval for when returns are filed
  - New Census Block Anomaly – Has Census population but no returns in prior year
- Additional Value Added (Discovered during development)
  - Out-of-State P.O. Box Frequency – By zipcode
  - Mail Forwarding Services – Identified service forwarding large amount of returns for “residents” of no income tax states
Revised Solution

- Interactive Drill Down Report That Grouped Similar Anomalous Returns Together
- Included All Components Of The Initial Solution, plus +
  - Analytical Model Anomaly (older model)
  - Analytical Model Geographic Anomaly – by block
  - Refund to Wage Anomaly – >25%
  - Income to Withholding Anomaly – >25%
  - Many others
Results Census

• Implemented 10/1/2014
• Not Used
Reasons for “TBD” Status

• Final Product not Ideal for Organizational Structure
  • End product required research of an analytical nature
  • After finding one anomalous return, would guide you to patterns
  • Did not “auto-suspend”

• Learning Curve
  • Should have used E-file Database
  • Daily Loads not Possible (anticipated at kickoff)
  • Current Year Data Loaded after Start of Filing Season
Successful Case: Fraud Scoring Algorithm
Business Case

• Agency Focus -- Taxpayer Service
  • Reduce false positives while getting legitimate taxpayers their money in a timely manner
  • Balancing taxpayer service while protecting State and residents
Proposed Solution

• Idea: Utilize scoring algorithm to “Triage” all returns

• Estimate probability return is fraudulent based off historical correlations

• Increase fraud $ while increasing efficiency
Implementation

• In 2011 deployed scoring algorithm directly in tax processing system
• Tried for three years with mixed results
• In 2015 re-estimated and deployed in data warehouse
• Results are preliminary but Very Promising
First Attempt

Perform Scoring in Processing System

First In Service For TY2011
Previously Identified Fraud

Historical Tax Returns

Estimate Coefficients

Example: Score = 0.3(FAGI) + 0.1(WH)

Current Year Tax Return

Extract Selected Data Elements

Tax Processing System

Score < 0.5
- Issue Refund

Score > 0.5
- Manual Review

Issue Refund

Manual Review
Results: Model #1

• Active for tax years 2011-2013
• 72,086 Returns ID’ed (24,000/Year)
• 10.53% Returns ID’ed were fraud
• $14.4 Million in fraud ($4.8M/Year)
Second Attempt:

Perform Scoring in Data Warehouse
Historical Tax Returns

Current Year Tax Return

Previously Identified Fraud

Extract all Fields of Return

Data Warehouse – Decision Tree Lives Here

Score < 0.5

Tax Processing System

Score > 0.5

Manual Review

Issue Refund
Results: Model #2

• Deployed on 4/6/2015
  • Two months active
  • Not Peak Fraud Time
• 5,869 Returns ID’ed (Est. 33K/Year)
• 55.31% Returns ID’ed were fraud
• $7 Million in fraud (Est. $24M/Year)
• Auto-suspends!
## Results: Model #2

**Comptroller of Maryland**  
**TRACE Analytics – Miscellaneous Fraud**

### Table 3: Tax Year 2013 Model Results

<table>
<thead>
<tr>
<th>Scores</th>
<th>Returns</th>
<th>Stopped By Legacy QRDT Processes</th>
<th>NOT Stopped By Legacy Processes</th>
<th>Known Fraud</th>
<th>Known Fraud Rate</th>
<th>Adjusted Fraud Rate</th>
<th>Percentage of Known Fraud Dollars Captured</th>
<th>Fraud Dollars Recovered (Millions)</th>
<th>Projected Unrecovered Fraud (Millions)</th>
<th>Total Dollar Potential (Millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;=0.80</td>
<td>13,260</td>
<td>12,167</td>
<td>1,093</td>
<td>9,377</td>
<td>71%</td>
<td>77%</td>
<td>71%</td>
<td>$16.4</td>
<td>$1.3</td>
<td>$17.7</td>
</tr>
<tr>
<td>&gt;=0.60</td>
<td>23,686</td>
<td>17,782</td>
<td>5,904</td>
<td>11,047</td>
<td>47%</td>
<td>62%</td>
<td>79%</td>
<td>$18.8</td>
<td>$5.3</td>
<td>$24.1</td>
</tr>
<tr>
<td>&gt;=0.5</td>
<td>31,057</td>
<td>21,393</td>
<td>9,664</td>
<td>11,569</td>
<td>37%</td>
<td>54%</td>
<td>82%</td>
<td>$19.4</td>
<td>$7.3</td>
<td>$26.7</td>
</tr>
<tr>
<td>Current Capacity</td>
<td>110,784</td>
<td>42,121</td>
<td>68,663</td>
<td>12,911</td>
<td>12%</td>
<td>31%</td>
<td>92%</td>
<td>$21.1</td>
<td>$24</td>
<td>$45.2</td>
</tr>
</tbody>
</table>
Output 1 - Worklist

**Comptroller of Maryland QRDT**
*Misc. Fraud Model Scores Less Than 1*

<table>
<thead>
<tr>
<th>SMART Trans ID</th>
<th>Suspense Code</th>
<th>Score Date</th>
<th>Index Score</th>
<th>MEF Extr Date</th>
<th>Form Year</th>
<th>Form Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>311515410048</td>
<td></td>
<td>8/4/15</td>
<td>0.9756412734487570</td>
<td>8/4/15</td>
<td>2014</td>
<td>MD02</td>
</tr>
<tr>
<td>311515400023</td>
<td></td>
<td>8/1/15</td>
<td>0.9756412734487570</td>
<td>5/31/15</td>
<td>2014</td>
<td>MD02</td>
</tr>
<tr>
<td>311515400073</td>
<td></td>
<td>8/1/15</td>
<td>0.9756412734487570</td>
<td>9/31/15</td>
<td>2014</td>
<td>MD02</td>
</tr>
<tr>
<td>311515400085</td>
<td></td>
<td>8/1/15</td>
<td>0.9756412734487570</td>
<td>9/1/15</td>
<td>2014</td>
<td>MD02</td>
</tr>
<tr>
<td>311515400008</td>
<td></td>
<td>8/1/15</td>
<td>0.9756412734487570</td>
<td>9/1/15</td>
<td>2014</td>
<td>MD02</td>
</tr>
<tr>
<td>311515400004</td>
<td></td>
<td>8/1/15</td>
<td>0.9756412734487570</td>
<td>9/1/15</td>
<td>2014</td>
<td>MD02</td>
</tr>
<tr>
<td>311515400051</td>
<td></td>
<td>8/1/15</td>
<td>0.9756412734487570</td>
<td>9/1/15</td>
<td>2014</td>
<td>MD02</td>
</tr>
<tr>
<td>311515400099</td>
<td></td>
<td>8/1/15</td>
<td>0.9756412734487570</td>
<td>9/1/15</td>
<td>2014</td>
<td>MD02</td>
</tr>
<tr>
<td>311515400105</td>
<td></td>
<td>8/1/15</td>
<td>0.9756412734487570</td>
<td>9/1/15</td>
<td>2014</td>
<td>MD02</td>
</tr>
<tr>
<td>311515400107</td>
<td></td>
<td>8/1/15</td>
<td>0.9756412734487570</td>
<td>9/1/15</td>
<td>2014</td>
<td>MD02</td>
</tr>
<tr>
<td>311515400112</td>
<td></td>
<td>8/1/15</td>
<td>0.9756412734487570</td>
<td>9/1/15</td>
<td>2014</td>
<td>MD02</td>
</tr>
<tr>
<td>311515400113</td>
<td></td>
<td>8/1/15</td>
<td>0.9756412734487570</td>
<td>9/1/15</td>
<td>2014</td>
<td>MD02</td>
</tr>
<tr>
<td>311515400115</td>
<td></td>
<td>8/1/15</td>
<td>0.9756412734487570</td>
<td>9/1/15</td>
<td>2014</td>
<td>MD02</td>
</tr>
<tr>
<td>311515400117</td>
<td></td>
<td>8/1/15</td>
<td>0.9756412734487570</td>
<td>9/1/15</td>
<td>2014</td>
<td>MD02</td>
</tr>
<tr>
<td>311515400119</td>
<td></td>
<td>8/1/15</td>
<td>0.9756412734487570</td>
<td>9/1/15</td>
<td>2014</td>
<td>MD02</td>
</tr>
</tbody>
</table>
Output 2a – In House Enhancements
Output 2b – In House Enhancements

![SAP GIS Table Example](image)

**Table: Comptroller of Maryland QRDT**

<table>
<thead>
<tr>
<th>Index Score</th>
<th>Count</th>
<th>Preparer PTIN</th>
<th>Count</th>
<th>Wages</th>
<th>Count</th>
<th>Federalagi</th>
<th>Count</th>
<th>Requested Refund</th>
<th>Count</th>
<th>Routing Number</th>
<th>Count</th>
<th>Routing Number</th>
<th>Account Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>1</td>
<td></td>
<td>33</td>
<td>310,037</td>
<td>33</td>
<td>310,037</td>
<td>33</td>
<td>2,228</td>
<td>6</td>
<td>124,303,120</td>
<td>20</td>
<td>31,101,160</td>
<td>31,101,160</td>
</tr>
<tr>
<td>0.8</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2,141</td>
<td>5</td>
<td>83,001,800</td>
<td></td>
<td>83,001,800</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>83,001,800</td>
<td></td>
<td>83,001,800</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>83,001,800</td>
<td></td>
<td>83,001,800</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>83,001,800</td>
<td></td>
<td>83,001,800</td>
<td></td>
</tr>
</tbody>
</table>
Output 2c – In House Enhancements
Annual Results Comparison
Legacy Scoring vs. Updated Scoring

Processing System:
- 24K Returns ID’ed
- 10.53% Positive ID
- $5M Refund Fraud

Data Warehouse:
- 33K Returns ID’ed
- 55.31% Positive ID
- $24M Refund Fraud
Annual Results Comparison
The BIG Picture

Legacy QRDT Processes

- 110K Reviewed
- ~ 10% Positive ID
- $23M Refund Fraud

Data Warehouse:

- 33K Returns ID’ed
- 55.31% Positive ID
- $24M Refund Fraud
QRDT Reaction

- Very Dedicated and **Passionate** Group
- Initial Reaction: Excitement with Apprehension
- "Shut Down" several legacy stand alone programs
  - Initial concern about letting some fraud through
  - With coaching, understood that yes, the model isn’t perfect, but with it you will stop more fraud!!!!
QRDT Reaction

Hi, Brian,

I was going to send this email for a while now but wanted to see more results from my research before I do that.

I know you got a copy of my email to Kathy before, where I was expressing my concerns about returns not being caught by preliminary model. But after new final model was put in place and I had a chance to do a little research almost every day for couple of weeks now, I can honestly say: YOUR SCORING MODEL ROCKS!!!!!

I did not keep the count of returns which I found through Data Warehouse, however I can say approximately about 90% of what I find, is being caught by your model (and probably bunch of others which I don’t find). It’s really very, very impressive what you have created! Hopefully, by next year, after we are done with all our projects, it will be catching even more.

I do see very few which are slipping through some cracks, I guess, and was wondering if you would like to look at them and may be figure out why they did not score higher.

Thank you so much for this AWESOME SCORING MODEL, Brian!

Best regards,

Alisa Charnley, QRDT
Data Warehouse Enhancements

• Use all fields from return in scoring model
• More advanced statistical procedures
  • Decision tree
• Data in same location as scoring allows for more dynamic model
  • Historical filing comparisons
• Adapt for fraud identified during year
Other Positive Contributing Factors

- Experience of Contractors/State Personnel
- Data Sources
- Desired End Product
- Processing System Constraints
Takeaway Points

• Fraud Analytics in DW can be Successful
• Success of one Single Project Uncertain
  • Final Product Characteristics
  • Data/Organization Differs from State to State
• Need Consistent Support from Management
• Lessons Learned from each Project Transferable to Subsequent Projects
Contact Info

- Andrew Schaufele
- Comptroller of Maryland
  - Director, Bureau of Revenue Estimates
  - 410.260.7450
  - aschaufele@comp.state.md.us