Payroll Audit Selection Model Redesign

Sharen Dhami, Sr. Research Analyst
Ernest Urbanovich, Sr. Research Analyst

May 2011

WORK SAFE BC
WORKING TO MAKE A DIFFERENCE

BIA
BUSINESS INFORMATION & ANALYSIS

About WorkSafeBC

- WorkSafeBC is an independent statutory agency.
- It serves 2.3 million workers and more than 200,000 employers throughout BC.
- Mandate: To work with workers and employers
  - To promote the prevention of workplace injury, illness and disease
  - To rehabilitate those who are injured, and assist with timely return to work
  - To provide fair compensation to replace lost wages for injured workers during their recovery
  - To ensure sound financial management for a viable workers' compensation system
- Funded through insurance premiums paid by employers and through investment returns.

Insurance Premiums

- WorkSafeBC collects insurance premiums from employers to cover the costs associated with work-related injuries and diseases, including health care, wage loss, rehabilitation, and administration.
- An employer's premium is based on the payroll of its workers.
- The amount an employer pays is determined by this formula:

$$Premium = \frac{(Base\ Rate\ \pm Experience\ Rating\ Adjustment) \times Payroll}{100}$$

Payroll Audits

- **Business Problem:**
  - WorkSafeBC has approximately 200,000 employers reporting payroll each year.
  - Premiums collected are based on payroll ***reported*** by the employers.
  - Statistical sampling has determined that a portion of premiums are lost each year as a result of payroll underreporting.
  - To minimize the revenue leakage due to payroll underreporting, the audit department conducts payroll audits each year.

- **Audit Selection Method:**
  - Since 1999, WorkSafeBC's Business Information & Analysis (BIA) department has been providing assistance to Audit Operations in identifying candidate employers for the annual payroll audit.
  - Between 1999 and 2002, employers were selected for payroll audit based solely on the size of their premiums and their industry.
  - In audit year 2003, a **scorecard model** was implemented and has since been in use.

Scorecard Model

- The scorecard is a rank-based quantitative tool built in MS Access that combines concepts, risk scoring and expert knowledge.
- The results produced by the scorecard effectively recovered a portion of the estimated leakage each year since implementation.
- However, in the most recent audit years, a much smaller portion of the leakage was recovered, the lowest amount since the model's implementation.
- There are a number of reasons for the recent decline, one of which appears to be the model's loss of effectiveness.

Objectives

1. To generate a list of candidate employers for payroll year 2011 based on a *new* predictive model.

2. To test and evaluate predictive analytics and data mining software tools.

Success Criteria

- **Model Success Criteria**
  - Better identify high yield employers for auditing
  - Recover a greater percentage of the estimated leakage

- **Software Success Criteria**
  - Can be used with our internal data sources
  - Can provide analysis that is needed and cannot be done by our current set of tools
  - Ease of use (can be used by the majority of our Sr. Research Analysts)
  - Sufficient support (technical/user)
  - Easy to deploy solutions

Data Variables

- **Target:**
  - Underreporting indicator: 0=less than $500 recovered, 1=$500+ recovered

- **Inputs:**
  - Years since last audit
  - Premium (Size)
  - Base rate
  - Industry
  - Previous years payroll underreporting indicator
  - Other

Methodology

- **Historical audit results (more than 60,000 audits) used to develop and evaluate multiple models**

- **Model Development:**
  - Three *new* predictive models were constructed using various software tools and data transformations:
    1. **Reg1:**
       - Logistic regression model using same variables and transformations inputted into the original scorecard (**SC***) model
    2. **Reg2:**
       - Logistic regression model using all identified variables as inputs
    3. **Tree:**
       - Decision tree model automatically created by SAS RPM (Rapid Predictive Modeler) advanced feature
       - RPM automatically created and evaluated 10 models based on various criteria.
    4. ***SC:**
       - Original scorecard model created in MS Access.

Methodology (cont'd)

- **Model Evaluation:**
  - Four different assessment methods were used to evaluate and compare the models

1. **Assessment Amount Recovered:** The total assessment amount recovered by the top employers as identified by each of the models
   - Evaluates overall effectiveness: Which model can recover the greatest amount of the leakage?

2. **Cumulative Gains Chart:** The gain in net assessment amount recovered with each additional employer audited
   - Measures the performance of the targeting model: Which model can most effectively identify high yield employers?

3. **Classification Matrix:** Classifies each model's ability to target candidate employers with respect to the actual results
   - Evaluates the model's accuracy: Which model has the highest rate of true positives & true negatives (lowest rate of misclassification)

4. **Model Overlap:** The percentage of employers targeted for auditing by two models that are common to both models
   - Compares the difference/similarities between the models: Are the models identifying the same or different candidate employers to audit?

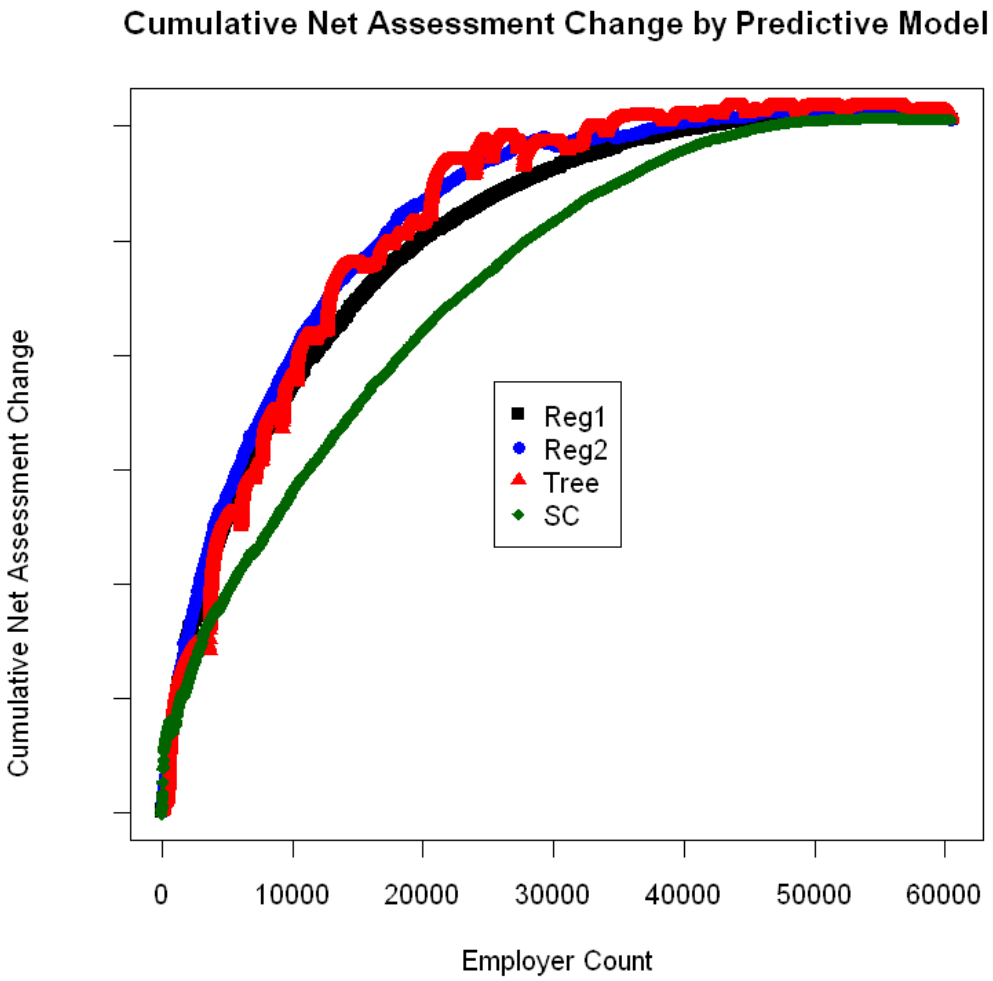
Results: Assessment Amount Recovered

Gain in Assessment Amount Recovered when compared to scorecard model (Baseline):

Model	Top 5,000 Employers
Reg1	35%
Reg2	45%
Tree	36%
Scorecard	0%

- The new models predict that a significantly greater amount would be recovered on the historical data than the scorecard model
  - Between 35-45% more
- However, the three new predictive models (Reg1, Reg2, Tree) predict similar recovery amounts

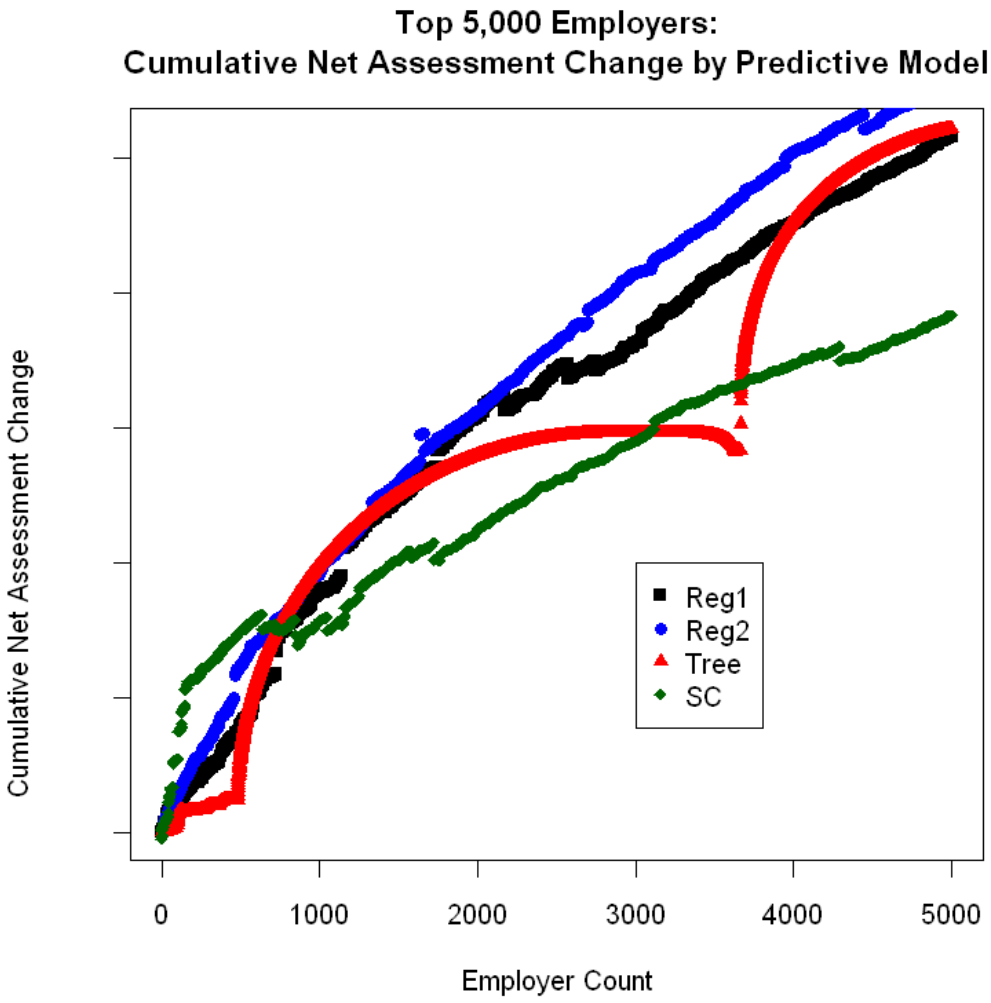
Results: Cumulative Gains Chart



- Based on historical data the three new predictive models are more effective than the score card model at identifying high yield employers
- Overall, Reg2 and the decision tree model are the most effective

BIA
BUSINESS INFORMATION & ANALYSIS

Results: Cumulative Gains Chart



- Although effective overall, the decision tree model is less effective than the regression models for identifying high yield employers when only the top 5,000 employers are considered
- Additionally, the tree model exhibits a "step" pattern, which may not be the most suitable given that the number of audits performed is constrained to the audit departments capacity

BIA
BUSINESS INFORMATION & ANALYSIS

Results: Combined Models

- The three models were combined based on their probability scores

Gain in Assessment Amount Recovered when compared to scorecard model (Baseline):

Model	Top 5,000 Employers
Reg1	35%
Reg2	45%
Tree	36%
Combined	49%
Scorecard	0%

- The combined model appears to show an improvement (4%) over Reg2
- Based on results and business needs the combined model was selected generate a list of candidate employers for Audit Year 2010

Challenges & Lessons Learned

- **Data Preparation**
  - Significant challenge
  - Merge data from various sources
  - Aggregate historical audit data at employer-year level
  - Point in time indicators
  - Transform data
  - We spent approximately 80% of the time on tasks related to data preparation
- **Understanding Business Processes**

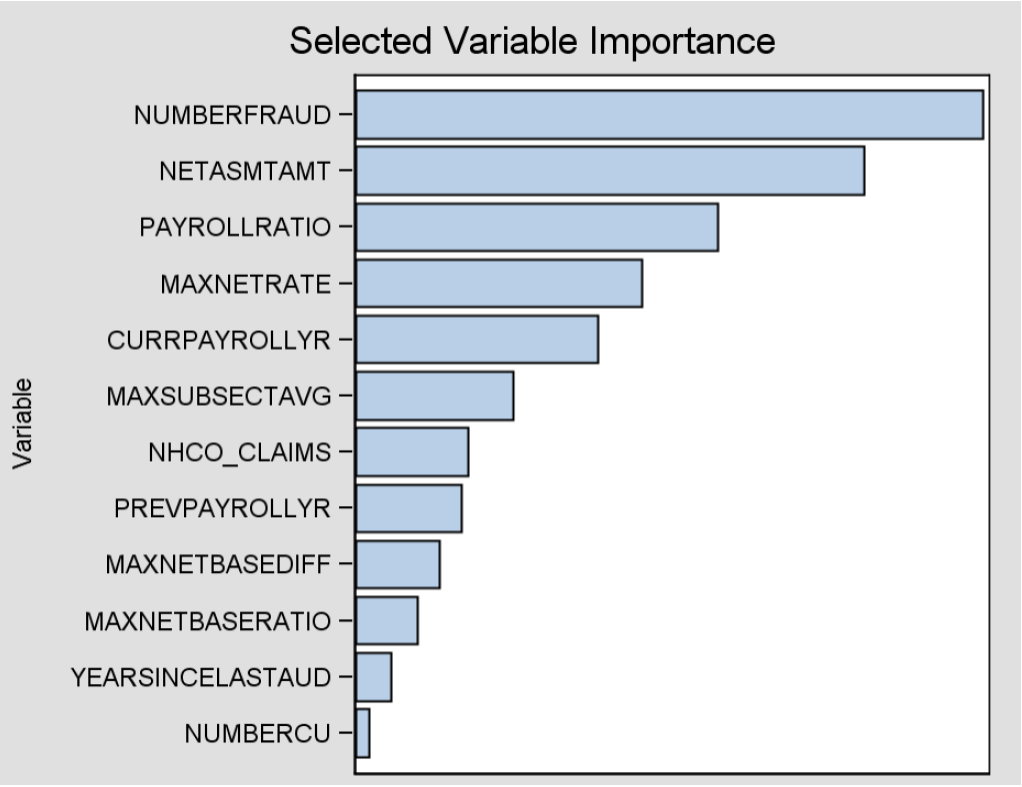
Summary & Conclusion

- The three predictive models created are a significant improvement over the scorecard model
- In the presence of a capacity constraint the regression models perform better than the decision tree model
- Transformations of the input data appear to have a significant impact on the results of the model
- The predictive power of the input variables appears to be limited
- Next Steps:
  - Identify additional input variables

Questions

Appendix: Importance of Variables

Tree Model



- The number of years the employer has had more than $500 recovered in past audits (NUMBER FRAUD) was the most significant factor, followed by the Net Assessment Amount and the ratio of the current years payroll to the previous years payroll (PAYROLLRATIO)
- Audit Officer and Audit Area were found to be insignificant predictors

BIA
BUSINESS INFORMATION & ANALYSIS

Results: Classification Matrix

Overall Classification

Actual	Reg1		Reg2		Tree	
	0	1	0	1	0	1
0	51%	20%	52%	20%	50%	21%
1	11%	18%	10%	19%	9%	20%
Correct Classification	69%		70%		70%	
Misclassification	31%		30%		30%	

- Overall, the three models have a similar rate of classification and misclassification

Classification of 0's and 1's

Actual	Reg1		Reg2		Tree	
	0	1	0	1	0	1
0	72%	28%	72%	28%	70%	30%
1	39%	61%	36%	64%	30%	70%

- The decision tree model most accurately identifies the target

Results: Model Overlap

Top 5,000 Employers

Model	Reg1	Reg2	Tree	SC
Reg1	100%	60%	39%	38%
Reg2	60%	100%	59%	32%
Tree	39%	59%	100%	20%
SC	38%	32%	20%	100%

- The overlap in the employers range between 20-60% for the top 5,000

- Although the assessment amounts recovered by the three models are similar, the employers that contribute to those amounts are quite different

  - A combined model may be more effective at recovering a greater portion of the assessment amount