Summary of the Data Mining Pilot Project

The data mining pilot project utilized the “data mart” developed by Information Services and attempted to demonstrate enhanced predictive capabilities using sophisticated data mining software. The project involved a steep learning curve. As could be expected, a number of technical problems were encountered. The time frame and data limitations precluded the possibility of approaching optimal predictive modeling. However, some positive if modest results were achieved.

Procedures and Results

The base for modeling was the known results of audits issued in Calendar Year 1997. Audit recovery was the dependent variable. The most predictive models developed from 1997 data were then used to predict 1998 audit results. All data manipulation and modeling was done with SPSS Clementine mining software. Data was obtained through ODBC connections to the SQL Server “data mart.” The data included:

- gross income, taxable income and tax due from the preceding 4 years.
- total deductions from the preceding 4 years.
- total wages and average # of employees from the preceding 4 years.
- 26 industry categories derived from SIC codes
- location (in-state vs. out-of-state)
- ownership type

In addition, flags were set on the presence or absence of line codes and deduction codes, and changes in variables from year to year were calculated. Unfortunately, the structure and limitations of the data made it impossible to utilize line code amounts, deduction type amounts and deduction line amounts. These variables may very well have provided the most predictive power with respect to audit recovery.

Only debit audits with no missing data were used in the modeling. Audit recovery amounts were grouped into four bands ($ 1 – 1,000, $1,000 – 5,000, $5000 – 10,000, over $10,000). “Rule induction” or “decision tree” models were built attempting to predict the recovery band for each audit. The 1997 audits were divided into a “training” sample and a “testing” sample. The software “learned” and built the model using the training data. The model was then used to predict the results of the test sample. This process tested the general applicability of the model. “ Overspecification” was a significant problem—the software had no problem deriving rulesets to accurately predict the training data, but the rulesets predictability did not generalize to the test data. The models which predicted the test 1997 data best were then applied to the 1998 data. The results from the best model are shown below.

Correct Band Predicted	52%
Prediction off by 1 Band	28%
Prediction off by 2 Bands	19%
Prediction off by 3 Bands	1%

Conclusions

Due to the limiting factors discussed above, the predictive power of the pilot model was positive but modest. As a learning experience, however, the pilot was an unquestioned success. A great deal of technical knowledge was acquired within the Department in a very short period of time. Some of the major lessons learned are as follows:

- Optimal data structures for query software are definitely not optimal for mining software—a "two-tiered" approach to data warehousing will frequently be necessary.

- The major part of data mining (possibly 85 to 95%) is data preparation and data cleansing.

- Optimal use of mining software requires "perfect" data, structured with fillers for missing records and missing fields.

- Despite the power of the modeling software, modeling is still a complicated process of structural design, analysis and experimentation.

- While training is essential and limited use of outside consultants may be beneficial, the Department does have the technical capacity to do data mining in-house.

The power of sophisticated mining software is truly phenomenal. At the same time, it is not a "magic bullet." Data mining requires a highly focused and structured approach. It is highly technical and resource intensive. However, for appropriate applications, sophisticated data mining could be an extremely valuable and cost effective strategy for the Department of Revenue.