# Predictive Analytics in the Public Sector: Using Data Mining to Assist Better Target Selection for Audit

**Duncan Cleary**
**Revenue Irish Tax and Customs, Ireland**
dcleary@revenue.ie

**Abstract**: Revenue, the Irish Tax and Customs Authority, has been developing the use of data mining techniques as part of a process of putting analytics at the core of its business processes. Recent data mining projects, which have been piloted successfully, have developed predictive models to assist in the better targeting of taxpayers for possible non-compliance/ tax evasion, and liquidation. The models aim, for example, to predict the likelihood of a case yielding in the event of an intervention, such as an audit. Evaluation cases have been worked in the field and the hit rate was approximately 75%. In addition, all audits completed by Revenue in the year after the models had been created were assessed using the model probability to yield score, and a significant correlation exists between the expected and actual outcome of the audits. The models are now being developed further, and are in full production in 2011. Critical factors for model success include rigorous statistical analyses, good data quality, software, teamwork, timing, resources and consistent case profiling/ treatments. The models are developed using SAS Enterprise Miner and SAS Enterprise Guide. This work is a good example of the applicability of tools developed for one purpose (e.g. Credit Scoring for Banking and Insurance) having multiple other potential applications. This paper shows how the application of advanced analytics can add value to the work of Tax and Customs authorities, by leveraging existing data in a robust and flexible way to reduce costs by better targeting cases for interventions. Analytics can thus greatly support the business to make better-informed decisions.

**Keywords:** tax; predictive analytics; data mining; public sector; Ireland

## 1. Introduction: Revenue and Ireland

The Irish Revenue Commissioners were established by statute in 1923. Their mission is to serve the community by fairly and efficiently collecting taxes and duties and implementing Customs controls. In broad terms Revenue's work includes:

- Assessing, collecting and managing taxes and duties that account for over 93% of Exchequer revenue;
- Administering the Customs regime for the control of imports and exports and collection of duties and levies on behalf of the EU;
- Working in co-operation with other state agencies in the fight against drugs and in other cross departmental initiatives;
- Carrying out agency work for other departments;
- Collection of PRSI (Pay Related Social Insurance) for the Department of Social Protection;
- Provision of policy advice on taxation issues.

Net total receipts in 2009 were EUR 33bn. In 2010 receipts were ~ EUR 31.5bn.

Increasingly, Revenue is applying advanced analytics in its business processes. One of these initiatives is applying predictive analytical techniques to assist in better case selection for audit. This paper describes this approach as applied recently by Revenue.

## 2. Revenue's research and analytics branch

Revenue's Research and Analytics Branch conducts program-wide and macro-level research at a corporate level. The branch conducts analyses to transform data into information often using SAS analytical software. The branch's work in Revenue includes large sample surveys, data mining (population profiling/ customer segmentation, pattern recognition, forecasting, predictive modelling) data quality exercises, experimental design for evidence based decision support, economic research and risk analysis. The branch uses both Revenue data and data from other sources. This work enables Revenue to make better use of its data and provides an improved understanding of the taxpayer population. The results are used to better target services to customers and to improve compliance.

This paper focuses on a number of recent uses of predictive analytics in Revenue.

Reference this paper as: Cleary, D. "Predictive Analytics in the Public Sector: Using Data Mining to Assist Better Target Selection for Audit" *Electronic Journal of e-Government* Volume 9 Issue 2 2011, (pp132 - 140), available online at www.ejeg.com

### 2.1 Business context

Revenue has a dual focus on its taxpayers. The first is on customer service, and the second is on compliance. RAB have conducted a number of exercises using analytics with a customer service focus (for example, see Clancy *et al*., 2010). This paper will focus the use of analytics from a compliance perspective, with the following types of targets:

- Likelihood to yield if audited
- Likely amount of yield if audited
- Likelihood to liquidate (business failure)

The aim of using analytics in Revenue is to show how analytics can assist the development of effective business strategies for Revenue, therefore optimising the use of Revenue resources. Analytics can demonstrably reduce costs, increase yields, and improve Revenue's service to taxpayers. Revenue is one of a number of tax authorities that employ analytics to improve business processes.

## 3. Why use predictive analytics for target selection?

Increasingly analytics are being used in companies and entities that are seeking means of making smarter decisions and getting better results by utilising their data assets, advances in computational power and software and a new emerging class of analysts who can extract the knowledge from the vast amounts of data and information currently available. Thought leaders in this area are offering sound guidance to those who wish to improve how they go about their business and achieve their goals. Revenue is pursuing the use of analytics as it recognises the value that can be gained by such an approach. The reader is referred to a number of recent publications for a more detailed exposition of why using analytics is making increasing sense in both private and public sector (Davenport & Harris, 2007; Davis *et al*., 2006; Miller *et al*., 2006; Davenport & Harris, 2010).

### 3.1 Data and variables: Data integration

The primary source of data for the predictive models is Revenue's sophisticated risk analysis programme, REAP (Risk Evaluation Analysis and Profiling), which electronically screens taxpayers' data covering several years. It uses ~300 business rules to quantify risk for approx. 800,000 taxpayer entities, and a risk run is created at least three times annually. Predictive analytics can be used to extend from the quantification of risk in a case, to predicting, for example, likelihood of yield, if a case were to be audited, and the potential amount of yield, or likelihood to liquidate. The inputs for predictive models were therefore the outputs from the REAP system.

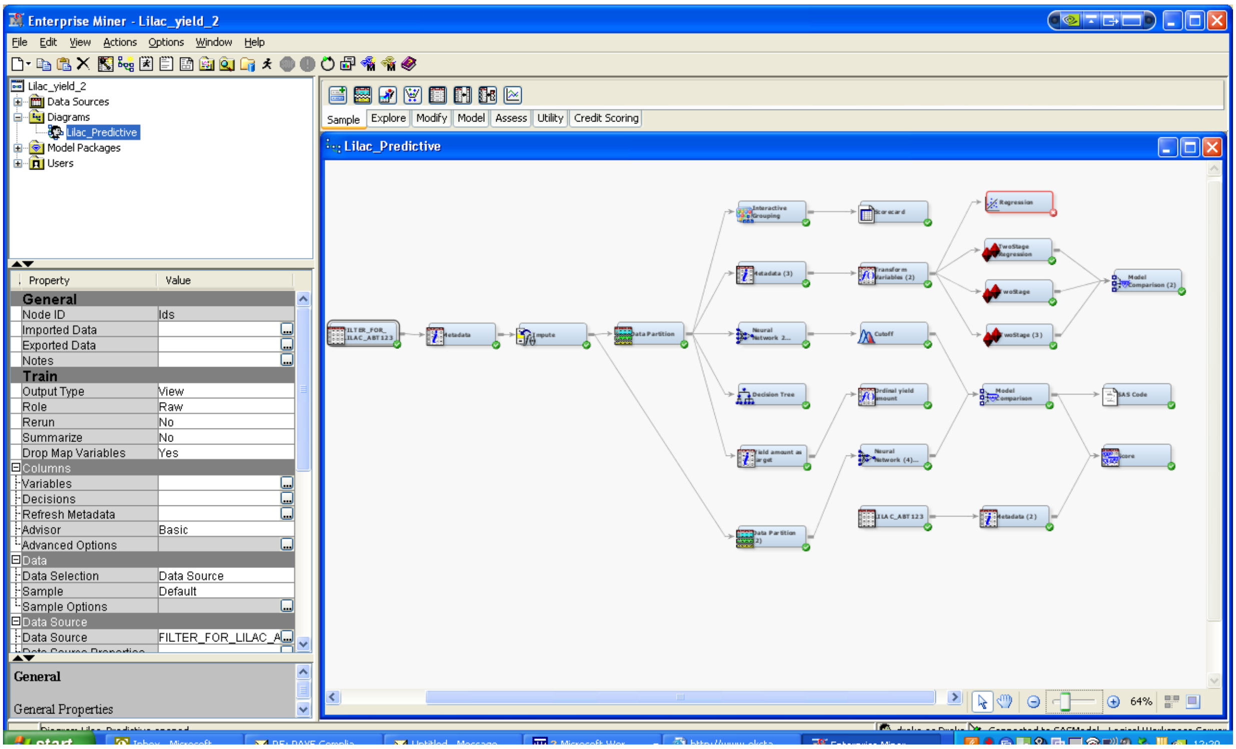
As with any Data Mining exercise, there is considerable effort required at the data integration stage, before modelling proper can begin. A process of ETL (Extraction, Transfer and Load) must be conducted, and RAB use SAS DI Studio (DI = Data Integration). The purpose of using this DI tool is to establish a process that would be scalable and semi automatic. Data from the REAP system and other sources in Revenue's Data Warehouse Environment are sourced and processed. Inherent in any data mining exercise is a review of data quality, which is not a trivial matter, and needs care and attention. The REAP system data is an opportunistic source of data, i.e. it is not designed specifically for predictive analytics. However, it is good quality in its format and completeness, which offers a solid platform for analysis. It is also readily available. Extensive work was required to understand the business context, the logic of the rules in REAP, and the underlying data that leads to those rules firing. A number of summarisations of the REAP system are produced for use in modelling, these included a table with the frequency of the rules fired in each case, and a binary indicator for rules firing/ not firing. In instances where data entailed many rows per entity, transpositions were performed, to create a flat file with one row per entity. Variables from the REAP system that summarise certain classes of rules, such as monetary risk and behaviour, are also assessed and incorporated for analysis. A target variable must be created for the training data (e.g. cases that had been audited in the previous two years). The target can be set as a binary target, e.g. where any yield over EUR 2,500= '1', and yield < EUR 2500 = '0'. The reason why the target should be set at EUR 2,500 is to avoid modelling for cases where the yield will be below the cost of a typical audit. If the target is the monetary amount of yield, this can be the second stage of a two stage model with the binary yield/ no yield as the first stage. In the case of liquidations, a database of known liquidations is used to train the model, using the profiles of the cases in REAP before the case liquidated. Additional data can include variables such as geography and economic sector.

This data integration process results in an ABT (Analytics Base Table) of one row per taxpayer with all of the attributes of interest as variables included for each taxpayer. These ABTs form the core inputs into the models. There will be a sub set where the target variable is populated (i.e. the data to be used to create the model). The bulk of the ABT will describe the population, which will be scored once the model is constructed. It should be noted that a lengthy exploratory analysis phase at univariate, bivariate and multivariate levels should be factored in and conducted with the raw input data and the final ABT. These univariate and bivariate analyses are performed by RAB in SAS Enterprise Guide (Ver. 4.3). Many variables are transformed to allow for better modelling, for example, to improve normality assumptions for continuous variables, or to optimally bin variables to better predict the target variable. 'Unsupervised' techniques (i.e. analyses without a target) are also conducted on the data, to further understand the data before predictive models are attempted. These unsupervised techniques can typically include cluster analysis and association analysis. If time allows, a full segmentation of the ABT can prove fruitful prior to modelling.

## 4. Data mining methodology

A predictive model by its nature has a target or interest. It can therefore be described as a 'Supervised' technique. Cases that have been audited (and been concluded), or cases that have gone into liquidation in the past can be used to train models. A predictive model produces a probability score for current and future cases of the likelihood of some outcome of interest occurring. This score can be deployed and used for decision support, e.g. case selection for audit. If the target is binary, both positive and negative outcomes need to be available to create a model to score unknown cases. If the target is numeric, e.g. monetary, a good range of values can be helpful.

RAB use SAS Enterprise Miner as their tool of choice for producing predictive models. Essentially the modelling process can be summarised as SEMMA: Sample, Explore, Modify, Model, & Assess. This process has been developed by SAS and it forms a solid framework for analysis. Scoring the full population completes the process. As cases are worked/ events happen, these can be used to improve the model. It is a very iterative process, with backwards and forwards movement within the SEMMA steps. Figure 1 illustrates a typical process flow.



| Property | Value |
|---|---|
| **General** | |
| Node ID | Ids |
| Imported Data | |
| Exported Data | |
| Notes | |
| **Train** | |
| Output Type | View |
| Role | Raw |
| Rerun | No |
| Summarize | No |
| Drop Map Variables | Yes |
| Columns | |
| Variables | |
| Decisions | |
| Refresh Metadata | |
| Advisor | Basic |
| Advanced Options | |
| Data | |
| Data Selection | Data Source |
| Sample | Default |
| Sample Options | |
| Data Source | |
| Data Source | FILTER_FOR_LILAC_A |

**Figure 1:** Screen grab of SAS E. Miner project, showing a SEMMA process flow

### 4.1 Sample

The ABT having been created, this essentially reflects a snapshot of the population of interest. Revenue is in the relatively luxurious position of not needing to worry about sampling as others may have to, who do not have access to full population data. However, training cases, or cases with a known outcome of interest, e.g. audit result, are often a biased sample. There are a number of ways of addressing this bias. One is to utilise a method known as 'Reject Inference'. This is a common method used in Banking and Insurance, wherein cases that have been accepted, for example for a loan, form a biased sample for modelling. Reject Inference assigns an imputed outcome to cases (based on their profile) that never had a loan, so that they can be used in further modelling. This method has been used successfully by Revenue in instances where audit cases, which form an analogous biased training dataset, can be augmented by cases as yet not audited.

Another means of addressing bias is to conduct random sample exercises, and record the outcome of interest for the random cases. This has two advantages. Firstly, it gives an approximation of the true proportions of, for example, a binary target's two levels in the population, which may be very different to the proportions in biased training data. These proportions can be used to adjust any model using biased training data by telling SAS E. Miner in the Decisions Processing function the approximation for the true population proportions. Secondly, a random sample may surface factors that lead to an outcome of interest, that are not typically used in case selection, thus giving a better picture of the overall population, for example the full range of taxpayer risk types. Revenue has therefore used data from its Random Audit program as inputs to training predictive models.

Over-sampling can be used when modelling rare events, such as liquidations. This is a process whereby when the target is binary (0,1), for example, a 100% representation of the '1's is included in the sample, and a complimentary random sample of '0's is selected from the population, often as a similar proportion. Thus one may have 50:50 proportioned '1's and '0's in the training data, even though the proportion of '1's in the population is very low, e.g. less than 1%. The purpose of this exercise is to avoid creating a model that is very accurate, but that classifies every case as a '0'.

Data partition also occurs at the sampling stage. This is where the training data is divided randomly into two or three groups. These groups are used for different aspects of the modelling process. In SAS they are labelled 'Training', 'Validation', and 'Test'. The proportions of cases in each category can vary according to the modeller's preferences and the data available. A model will be created using the training data, often judged by the validation data, and finally verified by the test data (which is not used in model creation). One must maximise the data available for modelling, while allowing enough data for validation. Typically proportions such as 60:20:20, 70:30:0 and 40:30:30 are used, depending on the model method used and the amount of training data available.

## 4.2 Explore

A thorough statistical exploratory analysis must take place next, before modelling. Univariate, bivariate and in some instances multivariate (perhaps 'unsupervised') analyses should be conducted. This can surface any data quality issues, data distribution and content should be assessed, outliers identified etc. The findings from this analysis may demand that the ABT be reconfigured, or that the raw data is augmented, or that new derived variables be created. This process then leads to the next stage of data modification.

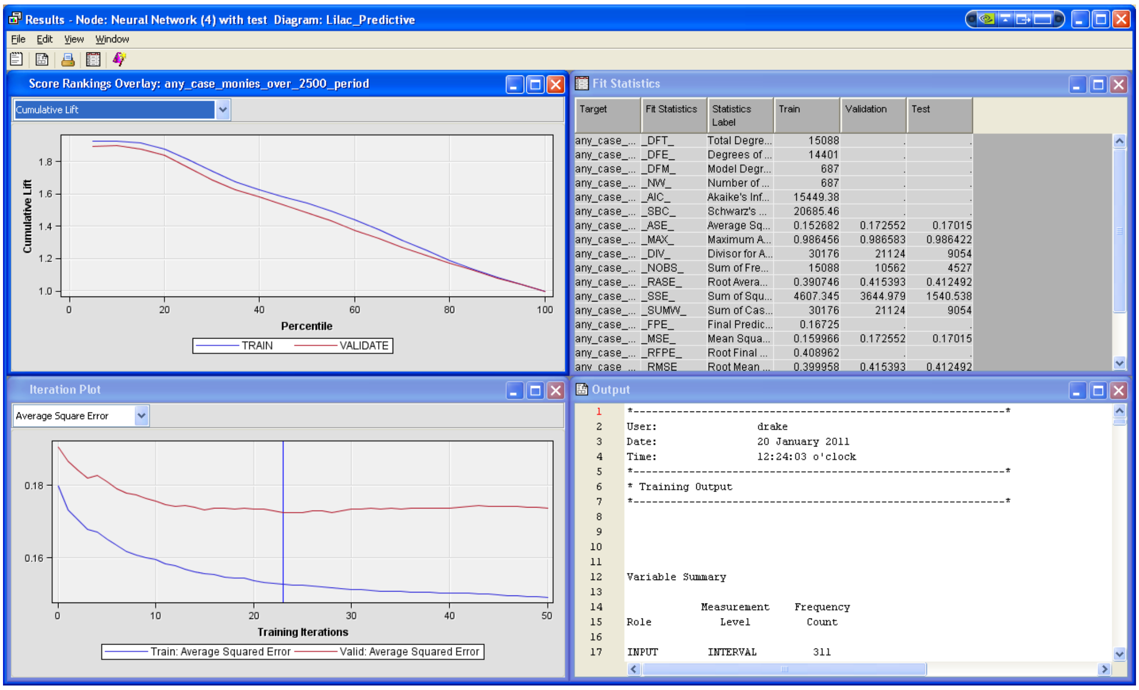
## 4.3 Modify

Nearly always in predictive modelling, it is wise to perform some transformations of the data. Typically this involves log (or other) transformations of highly skewed numeric variables, imputation of values where they are missing in the data, binning of categorical variables to remove rare levels, ranking of numeric data and creating a monotonic relationship between independent variables and the target variable. This last method is typically achieved by RAB using SAS Credit Scoring 'Interactive Grouping' node, which allows manual adjustment of the bins in each variable to maintain their usefulness at predicting the target, but to also ensure that they do so in a logical way.

## 4.4 Model

Once the process flow has been developed to a sufficient point, modelling can begin.

Several model methods are available in SAS E. Miner. Typically, a regression (for binary targets use logistic regression), a decision tree, and a neural network are chosen to prepare models. Various parameters can be set to suit the model requirements (see Sarma, 2007). For example, with logistic regression, a logit stepwise regression can be specified. The ranges of the options that are available in SAS E. Miner are beyond the scope of this paper, and the reader is referred to www.sas.com for details. Often, once models are produced, there is an iterative process involving returning to earlier steps in the model process flow. Issues heretofore unseen can be highlighted in the model results, and the modeller needs to adjust for problems like over-fitting. Typically many parameter changes will be made to maximise the predictive performance of the models, by for example making changes that increase the lift provided by the models, and by getting similar results for both training and validation data (i.e. making the model more robust). It should be noted however, that tweaking model parameters rarely exceeds the addition of new variables as a means of improving model performance.

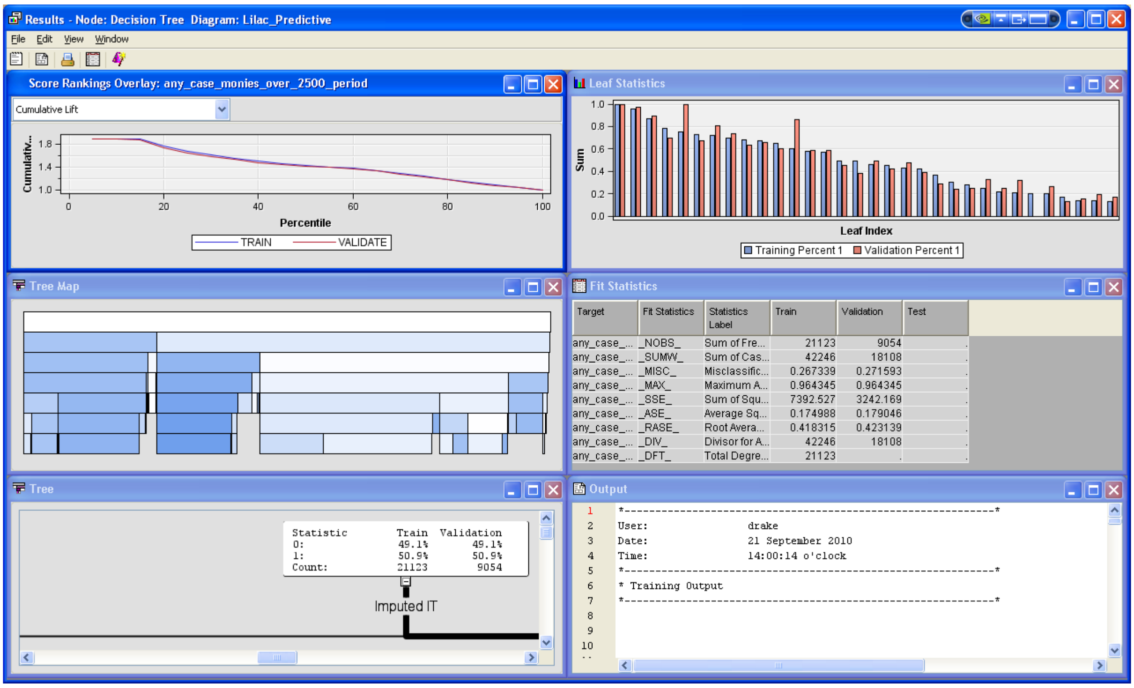


| Target | Fit Statistics | Statistics Label | Train | Validation | Test |
|---|---|---|---|---|---|
| any_case_... | _DFT_ | Total Degre... | 15088 | . | . |
| any_case_... | _DFE_ | Degrees of ... | 14401 | . | . |
| any_case_... | _DFM_ | Model Degr... | 687 | . | . |
| any_case_... | _NW_ | Number of ... | 687 | . | . |
| any_case_... | _AIC_ | Akaike's Inf... | 15449.38 | . | . |
| any_case_... | _SBC_ | Schwarz's ... | 20685.46 | . | . |
| any_case_... | _ASE_ | Average Sq... | 0.152682 | 0.172552 | 0.17015 |
| any_case_... | _MAX_ | Maximum A... | 0.986456 | 0.986583 | 0.986422 |
| any_case_... | _DIV_ | Divisor for A... | 30176 | 21124 | 9054 |
| any_case_... | _NOBS_ | Sum of Fre... | 15088 | 10562 | 4527 |
| any_case_... | _RASE_ | Root Avera... | 0.390746 | 0.415393 | 0.412492 |
| any_case_... | _SSE_ | Sum of Squ... | 4607.345 | 3644.979 | 1540.538 |
| any_case_... | _SUMW_ | Sum of Cas... | 30176 | 21124 | 9054 |
| any_case_... | _FPE_ | Final Predic... | 0.16725 | . | . |
| any_case_... | _MSE_ | Mean Squa... | 0.159966 | 0.172552 | 0.17015 |
| any_case_... | _RFPE_ | Root Final ... | 0.408962 | . | . |
| any_case_... | _RMSE_ | Root Mean ... | 0.399958 | 0.415393 | 0.412492 |

```
*------------------------------------------------------------*
User:                drake
Date:                20 January 2011
Time:                12:24:03 o'clock
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

            Measurement    Frequency
Role           Level         Count

INPUT        INTERVAL         311
```

**Figure 2:** Results screen from a neural network

The outputs from each model are examined and then their comparative performance is assessed. Outputs from a neural network (Figure 2) and a decision tree (Figure 3) are shown as examples here.



| Target | Fit Statistics | Statistics Label | Train | Validation | Test |
|---|---|---|---|---|---|
| any_case_... | _NOBS_ | Sum of Fre... | 21123 | 9054 | . |
| any_case_... | _SUMW_ | Sum of Cas... | 42246 | 18108 | . |
| any_case_... | _MISC_ | Misclassific... | 0.267339 | 0.271593 | . |
| any_case_... | _MAX_ | Maximum A... | 0.964345 | 0.964345 | . |
| any_case_... | _SSE_ | Sum of Squ... | 7392.527 | 3242.169 | . |
| any_case_... | _ASE_ | Average Sq... | 0.174988 | 0.179046 | . |
| any_case_... | _RASE_ | Root Avera... | 0.418315 | 0.423139 | . |
| any_case_... | _DIV_ | Divisor for A... | 42246 | 18108 | . |
| any_case_... | _DFT_ | Total Degre... | 21123 | . | . |

```
Statistic          Train  Validation
0:                 49.1%       49.1%
1:                 50.9%       50.9%
Count:             21123        9054
```

```
*------------------------------------------------------------*
User:                drake
Date:                21 September 2010
Time:                14:00:14 o'clock
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*
```

**Figure 3:** Results screen from a decision tree

## 4.5 Assessment

Assessing models can be done in a number of ways. Three main methods are:

- Within model process-flow assessment
- Sample tests of cases ranked by a model
- Back validation of model with subsequent events.

This section covers the first of these three; the next two are covered under the Results section below. SAS E. Miner offers an assessment node. There are many criteria for assessing the performance of a model. As an example, here a logistic regression with a binary target will be used. Criteria such as misclassification or average square error can be used, but often the most effective measure of the success of a model is the ROC curve (Receiver Operating Characteristic). A ROC curve shows the values of the true positive fraction and the false positive fraction at different cut-off values (Sarma, 2007). The cut-off values can be set to maximise the number of true positives in a set of cases. A set of ROC charts can be produced, that compares Training, Validation and Test data (see Figure 4), for all of the models fed into the assessment node. Essentially, the model that looks most similar across the three modelling data sets, and also has the maximum distance between the curve and the diagonal line (which represents no predictive power, i.e. an equivalent to flipping a coin), is usually the best model to use for scoring the full population/ new cases.

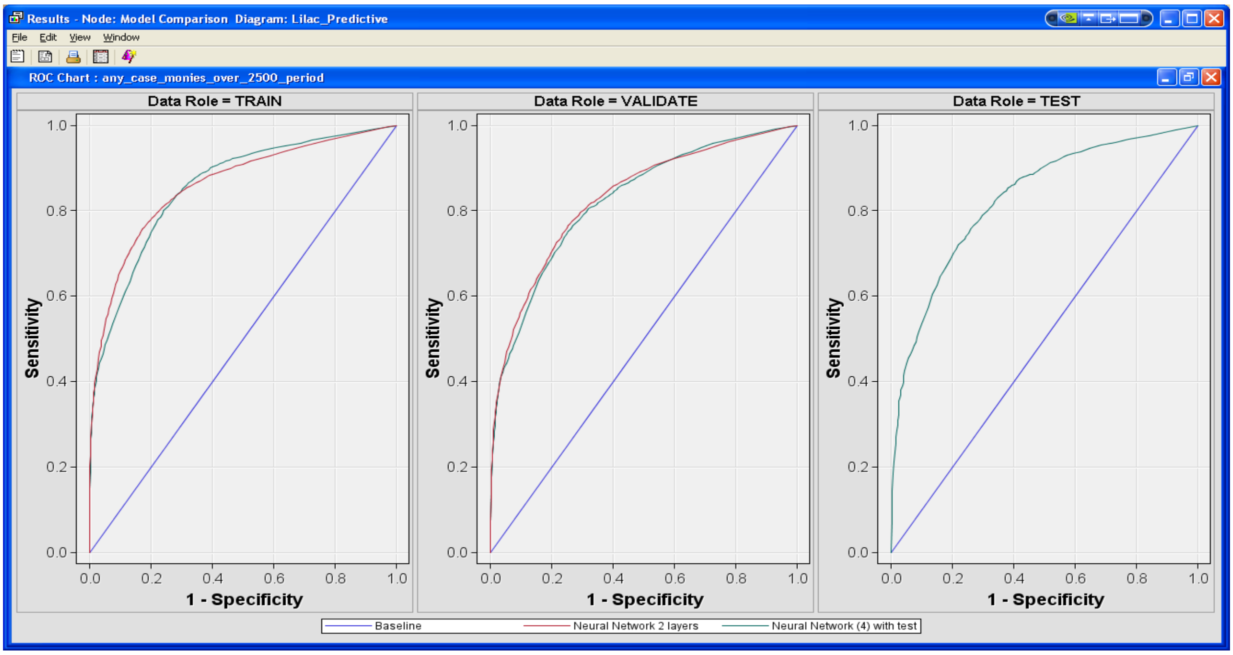


**Figure 4:** Results screen from a model comparison node

## 5. Results and evaluation

A number of successful models have been produced by RAB. Following within modelling process validation, these have also been field tested and back validated with success.

Each model has been used to score the population of relevance. These scores have been assessed, binned into deciles and demi-deciles, and have been used to rank the case base in descending order of likelihood. Cut-offs have been set based on criteria such as misclassification rate and the amount of cases that can be worked based on resources available.

For this paper the example results to be shown are for the Yield model, i.e. the Likelihood to yield if audited model.

## 6. Yield prediction model

A successful pilot exercise testing a predictive model focused on yield was conducted in 2009/2010. The pilot exercise resulted in a 3:1 hit rate for the cases provided for working. Both case specific reviews and 'Back Validation' against all closed audits suggested that the approach was robust.

Following on from this, a new predictive model was created using similar Data Mining techniques (the pilot used Credit Scoring, the subsequent model used a Neural Network model). The model uses data from the most recent REAP Risk run available as inputs, and was trained with closed audits from prior years, as before. As noted above, the purpose of the model is to identify cases, which have a similar profile to known yielding audits, and to rank the likelihood of yield in the event of an audit of those cases. The full case base receives a probability score, and for the purposes of this model a cut-off of 80% probability to yield has been set. This cut-off, which in effect creates a binary Yes/No indicator for each case, has been provided to auditors for case selection, through the REAP system. As such, each case selector and worker can access some of the power of this predictive analytics model. Approximately 40,000 cases (5% of case base) fired the rule. As with any model, feedback is critical for evaluation and model improvement.

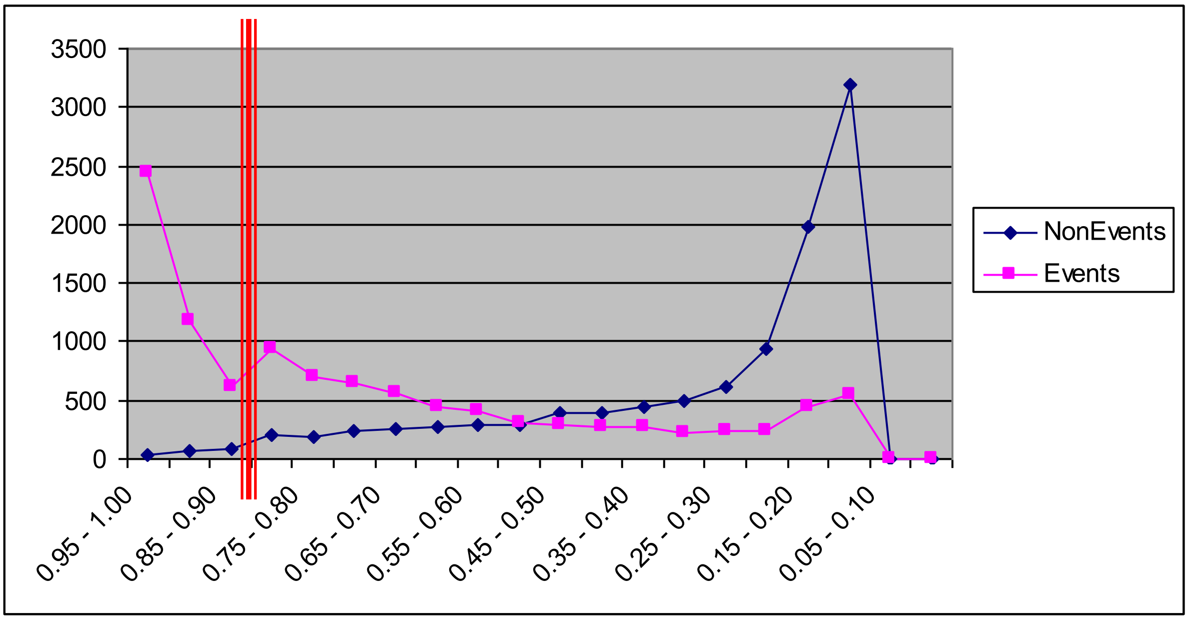


**Figure 5**: Training cases, with yielding and non-yielding, by model probability score. X axis: Probability to yield left to right, high to low, Y axis: Training cases, frequency, events= yielding cases, non events = non yielding

At the 0.8 (80%) cut-off shown by the treble red line, and above (i.e. to the left) in Figure 5, the majority of cases were yielding in training data. Thus the focus was placed on cases scoring similarly to these cases in the full risk run population.

An evaluation of the Predictive model using the latest REAP run was conducted in mid January 2011. The method used was 'Back Validation' of the 2010 yield prediction model, with closed audits (i.e. known results) from 2010. These audits were not used in the training of the model; hence the model can be assessed on the basis of how well it predicted events unknown to it at the time of its creation. The results were very positive, showing a strong relationship between predicted likelihood to yield and actual average yield. Each point on the graph below represents approximately 475 cases (total of 9500 cases represented in Figure 6). There is a strong positive correlation between realised yield and predicted audit outcome, i.e. as the probability to yield increases so too does the average (median) yield. This suggests that the model is robust and a good case selection tool. Cases from the uppermost probability deciles have yield on average twice that of cases from the lowermost deciles (Figure 7). Subsequent back validations continue to show that the model is performing well.

## 7. Conclusions

The Yield model has since been rerun a number of times (most recently the late 2011 REAP run), with modifications based on feedback, and incorporated in the nationally disseminated current REAP run in the form of an indicator rule for cases where there is a high probability to yield. In effect, the model output is available to case selectors. This qualifies the model as being in production. As cases are selected using this rule, results will be assessed and incorporated into future models. In addition, a liquidation probability model has also been made available by similar means, having been validated with events since model creation.

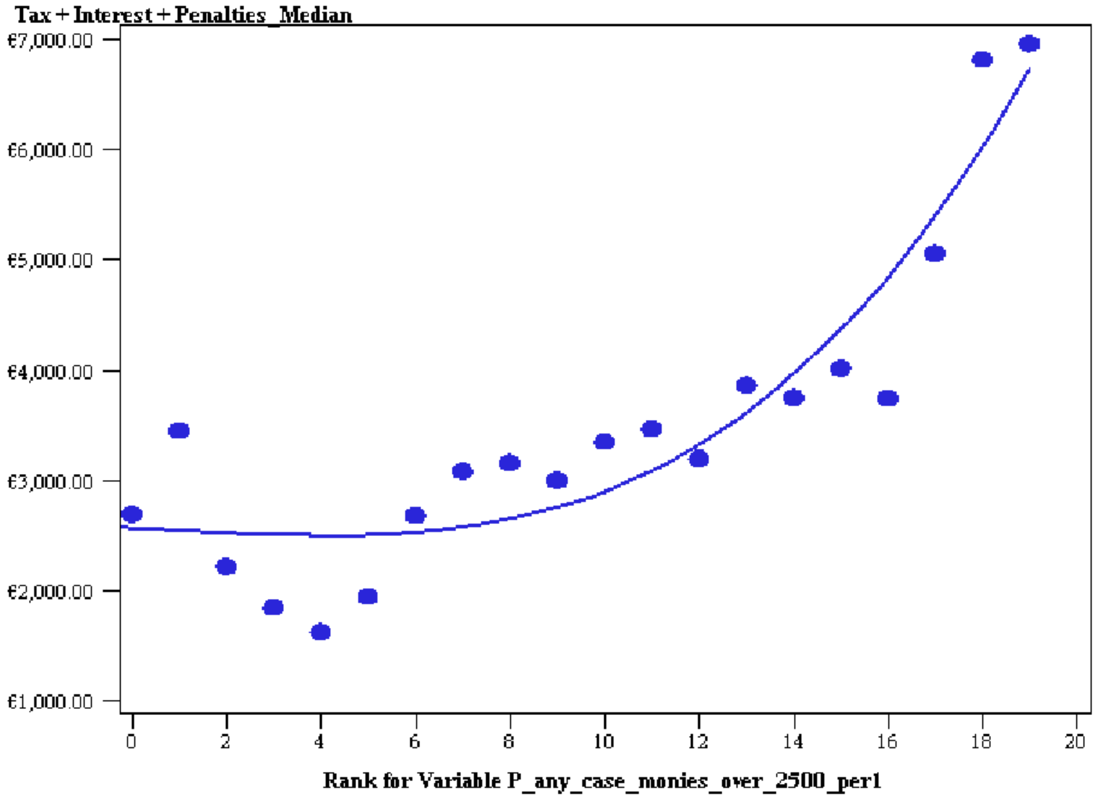


**Figure 6:** Correlation between modelled likelihood to yield and actual audit yield, probability demi-deciles and median yield per decile

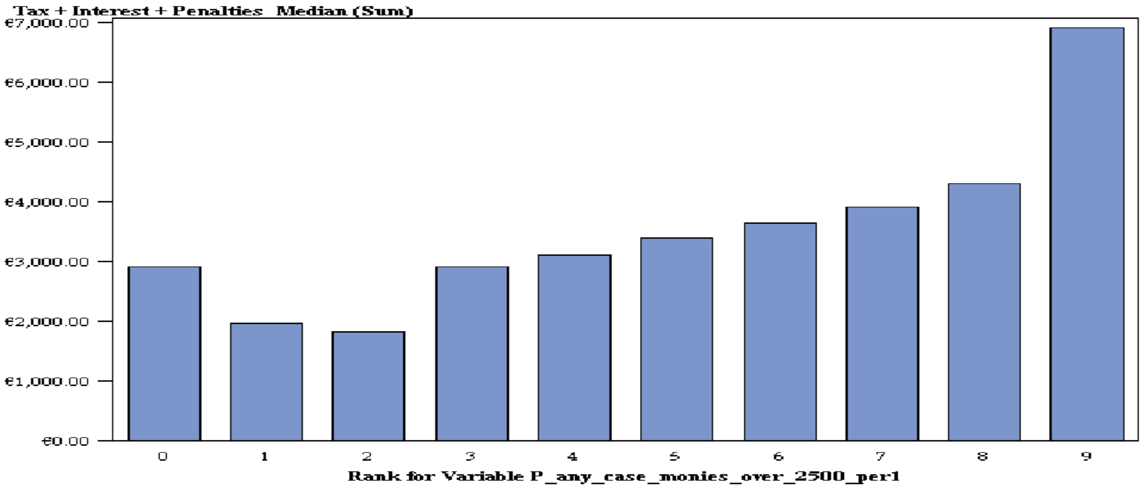


**Figure 7:** Median yield per decile of model probability to yield

RAB is developing more models for the business (e.g. quantity of monetary yield, using a two-stage model, models for specific economic sectors, regional models etc.). RAB continues to evaluate models through field-testing, in co-operation with Revenue regions. RAB thus hopes to extract more value from the data and information Revenue already has, and is increasingly making use of the power of analytics, and is making analytics more central to how Revenue performs its work.

## References

Clancy, J, Manai, G and Cleary, D. 2010. Segmentation of the PAYE Anytime Users Electronic Journal of e-Government Volume 8 Issue 2 2010, (pp105-120), available online at www.ejeg.com

Davenport, T.H. & Harris, J. 2007. Competing on Analytics: The New Science of Winning. Harvard Business School Press.

Davenport, T.H. & Harris, J. 2010. Analytics at Work: Smarter Decisions, Better Results. Harvard Business Press.

Davis, J., Miller, G.J. & Russell, A., 2006. Information Revolution: Using the Information Revolution Model to Grow Your Business. Wiley.

Miller, G.J, Bräutigan, D. & Gerlach, S.V. 2006. Business Intelligence Competency Centres: A Team Approach to Maximising Competitive Advantage. Wiley.

Sarma, Kattamuri S. 2007. Predictive Modelling with SAS Enterprise Miner: Practical Solutions for Business Applications. Cary, NC: SAS Institute Inc.

SAS E. Miner reference (accessed March 2011): http://www.sas.com/technologies/analytics/datamining/miner/index.html