
# NAVAL POSTGRADUATE SCHOOL

**MONTEREY, CALIFORNIA**

# THESIS

**DETECTION OF ERRONEOUS PAYMENTS UTILIZING SUPERVISED AND UNSUPERVISED DATA MINING TECHNIQUES**

by

Todd E. Yanik

September 2004

| | |
|---|---|
| Thesis Advisor: | Samuel E. Buttrey |
| Second Reader: | Lyn R. Whitaker |

**Approved for public release; distribution is unlimited**

# *Amateurs discuss strategy, Professionals study logistics*

| REPORT DOCUMENTATION PAGE | | Form Approved OMB No. 0704-0188 |
|---|---|---|



| 1. AGENCY USE ONLY *(Leave blank)* | 2. REPORT DATE | 3. REPORT TYPE AND DATES COVERED |
|---|---|---|
| | September 2004 | Master's Thesis |

| 4. TITLE AND SUBTITLE: Detection Of Erroneous Payments Utilizing Supervised And Unsupervised Data Mining Techniques | 5. FUNDING NUMBERS |
|---|---|
| 6. AUTHOR(S) Todd E. Yanik | |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)<br>Naval Postgraduate School<br>Monterey, CA 93943-5000 | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES)<br>Defense Finance and Accounting Service<br>Internal Review Seaside (Operation Mongoose) | 10. SPONSORING/MONITORING AGENCY REPORT NUMBER |

**11. SUPPLEMENTARY NOTES** The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

| 12a. DISTRIBUTION / AVAILABILITY STATEMENT | 12b. DISTRIBUTION CODE |
|---|---|
| Approved for public release; distribution is unlimited | A |

**13. ABSTRACT (maximum 200 words)**

In this thesis we develop a procedure for detecting erroneous payments in the Defense Finance Accounting Service, Internal Review's (DFAS IR) Knowledge Base Of Erroneous Payments (KBOEP), with the use of supervised (Logistic Regression) and unsupervised (Classification and Regression Trees (C&RT)) modeling algorithms. S-Plus software was used to construct a supervised model of vendor payment data using Logistic Regression, along with the Hosmer-Lemeshow Test, for testing the predictive ability of the model. The Clementine Data Mining software was used to construct both supervised and unsupervised model of vendor payment data using Logistic Regression and C&RT algorithms. The Logistic Regression algorithm, in Clementine, generated a model with predictive probabilities, which were compared against the C&RT algorithm. In addition to comparing the predictive probabilities, Receiver Operating Characteristic (ROC) curves were generated for both models to determine which model provided the best results for a Coincidence Matrix's True Positive, True Negative, False Positive and False Negative Fractions. The best modeling technique was C&RT and was given to DFAS IR to assist in reducing the manual record selection process currently being used. A recommended ruleset was provided, along with a detailed explanation of the algorithm selection process.

| 14. SUBJECT TERMS | 15. NUMBER OF PAGES |
|---|---|
| Data Mining, Erroneous Payments, Logistic Regression, Hosmer Lemeshow Test, Classification and Regression Trees, Receiver Operator Characteristic curves, supervised and unsupervised modeling. | 94 |
| | 16. PRICE CODE |

| 17. SECURITY CLASSIFICATION OF REPORT | 18. SECURITY CLASSIFICATION OF THIS PAGE | 19. SECURITY CLASSIFICATION OF ABSTRACT | 20. LIMITATION OF ABSTRACT |
|---|---|---|---|
| Unclassified | Unclassified | Unclassified | UL |

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

**Approved for public release; distribution is unlimited**

# DETECTION OF ERRONEOUS PAYMENTS UTILIZING SUPERVISED AND UNSUPERVISED DATA MINING TECHNIQUES

Todd E. Yanik
Lieutenant Commander, United States Navy
B.S., West Virginia University, 1990

Submitted in partial fulfillment of the
requirements for the degree of

**MASTER OF SCIENCE IN OPERATIONS RESEARCH**

from the

**NAVAL POSTGRADUATE SCHOOL**
**SEPTEMBER 2004**

Author: Todd E. Yanik

Approved by: Samuel E. Buttrey
Thesis Advisor

Lyn R. Whitaker
Second Reader

James N. Eagle
Chairman, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

# ABSTRACT

In this thesis we develop a procedure for detecting erroneous payments in the Defense Finance Accounting Service, Internal Review's (DFAS IR) Knowledge Base Of Erroneous Payments (KBOEP), with the use of supervised (Logistic Regression) and unsupervised (Classification and Regression Trees (C&RT)) modeling algorithms. S-Plus software was used to construct a supervised model of vendor payment data using Logistic Regression, along with the Hosmer-Lemeshow Test, for testing the predictive ability of the model. The Clementine Data Mining software was used to construct both supervised and unsupervised modeling of vendor payment data using Logistic Regression and C&RT algorithms. The Logistic Regression algorithm, in Clementine, generated a model with predictive probabilities, which were compared against the C&RT algorithm. In addition to comparing the predictive probabilities, Receiver Operating Characteristic (ROC) curves were generated for both models to determine which model provided the best results for a Coincidence Matrix's True Positive, True Negative, False Positive and False Negative Fractions. The best modeling technique was C&RT and was given to DFAS IR to assist in reducing the manual record selection process currently being used. A recommended ruleset was provided, along with a detailed explanation of the algorithm selection process.

THIS PAGE INTENTIONALLY LEFT BLANK

# TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

# LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

# LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

# ACKNOWLEDGMENTS

I would like to thank to Professor Samuel E. Buttrey and Professor Lyn R. Whitaker for their guidance and continuous support in carrying out this thesis research. I would also like to thank Lieutenant Colonel Chris Nelson and Mr. Dave Riney for their assistance and the opportunity to work on a project at Operations Mongoose, Defense Finance Accounting System, Seaside, California. Their extensive knowledge of Operation Mongoose was quite helpful in completing this thesis.

I would like to thank all of my friends and colleagues for their camaraderie and good humor throughout my stay in Monterey. Without this group, the past two years would have been assuredly longer and certainly not as enjoyable.

Most importantly, I would like to thank my wife and daughters for their support. Cristina's patience and understanding during this time at Monterey was the guiding light and pillar of strength in achieving a Master's degree at the Naval Postgraduate School.

THIS PAGE INTENTIONALLY LEFT BLANK

# EXECUTIVE SUMMARY

The purpose of this thesis is to enhance current auditing techniques for detecting erroneous and fraudulent payments at Defense Finance Accounting Service (DFAS) payment activities. DFAS, Internal Review (IR) has been utilizing data mining techniques since 1999 to determine fraudulent and erroneous payments. Since 2002, DFAS IR has built a Knowledge Base of Erroneous Payments (KBOEP) during their continual auditing process of DFAS payment activities.

The record selection process is done by a mix of electronic and manual means. Records are electronically selected form one of five fraud detection models and then the records are sorted and viewed in the Microsoft Access database. Auditor experience and historical trends of erroneous payments provide much of the impetus for selecting records to be audited in the field. Due to the enormous number of records needing auditing, a push towards a more independent, statistical base record selection process should be developed to uncover trends that human intervention cannot detect. Models that can be developed and validated to select records for audit would allow the auditors to recoup more funds from erroneous payments.

The goal of this thesis was to look at the KBOEP and to apply supervised (Logistic Regression) and unsupervised (Classification and Regression Trees (C&RT)) algorithms to generate models to perform electronic record selection. Various tests, such as the Hosmer-Lemeshow Test and the Receiver Operating Characteristic (ROC) curves, will be

used to determine the strength of one technique over the other. The research presented in this thesis will allow the DFAS IR auditors to improve an already outstanding operation. These techniques should allow the auditors to focus more on trends that have not been captured previously so as to further reduce the loss of funds through erroneous and fraudulent behavior by government and civilian payment activities.

# I. INTRODUCTION

## A. PURPOSE

The purpose of this thesis is to enhance current auditing techniques for detecting erroneous and fraudulent payments at Defense Finance Accounting Service (DFAS) payment activities. These efforts are lead by DFAS, Internal Review (IR) Seaside, also known as Operation Mongoose. DFAS IR has been utilizing data mining techniques since 1999 to determine fraudulent and erroneous payments. Since 2002, DFAS IR has developed a database of known erroneous payments during their continual auditing process of payment activities. This database is examined to improve the current auditing process used by the DFAS auditors.

## B. BACKGROUND

The IR Seaside Office was put into operation to detect fraudulent payments within the Department of Defense's (DOD) payment activities. The investigative team's original name was Operation Mongoose. It was noted during an analysis conducted by Oxendine [1] that numerous fraud problems were occurring within the DOD in the mid-1990's.

Operation Mongoose utilized new data mining technology to search the vast numbers of DFAS vendor pay transactions for potentially fraudulent payments. Of these payments, sixteen were prosecuted successfully for fraud. The Operation Mongoose team determined from its knowledge base that there were four fraud types that could be used as a foundation for building supervised classification and prediction models. DFAS IR absorbed Operation Mongoose in 1999, along with the work of conducting audits of payment

records in search of Conditions Needing Improvement (CNI), overpayments, detection of duplicate payments and potentially fraudulent behavior. IR Seaside does not conduct the audits, but assists the auditors in understanding the data mining techniques and collects audit results for future analysis.

## C. RESEARCH GOALS

This paper will enhance existing manual auditing techniques and identify areas for improvement. The development of models that can search field payment records electronically will aid in reducing the manual process of searching through a database for erroneous or fraudulent payments, allowing the auditors to focus more on the audits themselves. In this thesis, we will look at which statistical analysis technique will work best in developing and enhancing existing erroneous payment models.

Chapter I and II provide an introduction and background to DFAS IR's operation. In Chapter III, we include a brief review of the data mining techniques used by DFAS IR and how they were developed. This review includes both supervised and unsupervised modeling techniques and a review of previous site audits conducted by DFAS IR for 2002 and 2003. In Chapter IV, we include a thorough review of the techniques to be used. These techniques are Logistic Regression, Classification Trees, Hosmer-Lemeshow Test and Receiver Operating Characteristic curves. These tools help develop the predictive models for improving the manual audit selection process for one of five DFAS IR fraud models. In Chapter V, a thorough review of the results of the analysis is performed, a statistical technique is chosen and a model is recommended for

selecting erroneous payments electronically. Finally in Chapter VI, conclusions and recommendations are provided.

THIS PAGE INTENTIONALLY LEFT BLANK

# II. LITERATURE REVIEW

As noted in the General Accounting Office Report, GAO-02-069G, *Strategies to Manage Improper Payments: Learning from Public and Private Sector Organizations,* improper payments are a widespread and significant problem in the federal government and among states, foreign governments, and private sector companies. While in the private sector improper payments most often present an internal problem that threatens profitability, in the public sector they can translate into serving fewer recipients or represent wasteful spending or a higher relative tax burden. These wasteful and erroneous payments prevent taxpayer resources from meeting the missions and goals of the Department of Defense. The root causes of improper payments can normally be tracked to a lack of, or a breakdown in, internal controls. The risk of improper payments increases in programs with complex criteria for computing payments, a significant volume of transactions, or emphasis on expediting payments. [4]

Internal controls are not one event, but a series of actions and activities that occur throughout companies operations. People make internal controls work, and responsibility for good internal controls rests with all managers. [5] No matter how all-inclusive the auditing techniques are, they require knowledgeable, diligent and ethical people to perform the auditing process.

Fraudulent and erroneous payments are a problem not only in the DOD, but in the civilian sector as well. It is important to understand that too close relationships between companies and auditors are occurring today in the

civilian sector and having an impact on the integrity of the auditing process. The Arthur Anderson and ENRON auditing scandal exemplifies this problem of poor audit controls. In a proportion common to the Big Five accounting firms, half the $52 million a year Arthur Andersen collected from ENRON was for its accounting services, and the other half was for its consulting business. [17] As these relationships occur it is important to develop auditing techniques that will attempt to detect trends in large volumes of data. As payments are made electronically, the need to analyze and detect irregularities in those payments increases. The lack of consistent and independent audits among large corporations may lead to fraudulent behavior not being kept in check or detected. The audit process for civilian and public sector financial transactions should be kept separate and distinct from the activity under audit in order to maintain financial integrity.

The United States General Accounting Office (GAO) conducted a study of the techniques used by government agencies to detect or prevent fraud and improper payments. [4] The study cites a number of activities that are using data mining to detect abnormalities. For instance, the Illinois Department of Public Aid applies data mining techniques to detect fraudulent billing and kickback schemes. Another case reveals how the Texas Health and Human Services Commission is using neural networks to identify fraudulent claims. The Texas commission successfully identified over six million dollars for

recovery in fiscal year 2000. The GAO also reports on a number of other institutions and the data mining techniques used in fraud detection efforts. [2]

Several theses at the Naval Postgraduate School have looked at improving the audit process for DFAS IR. Jenkins [2] looked at DFAS IR's use of data mining techniques to analyze millions of vendor transactions each year in an effort to combat fraud. The long timeline required to investigate potential fraud precludes DFAS from using fraud as a supervised modeling performance measure, so instead it uses the conditions needing improvement (CNI) found during site audits. The research evaluated supervised models to determine if models improved with each new audit and proposed four initiatives to enhance the modeling process: a revised model scoring implementation, a knowledge base of audit results, alternative model streams for record selection and a recommended modeling process for the CNI knowledge base.

Roulliard [8] proposed a standardized procedure for detecting fraud in DFAS vendor payment transactions through unsupervised modeling (cluster analysis). Clementine Data Mining software was used to construct unsupervised models of vendor payment data using the K-Means, Two Step, and Kohonen algorithms. Cluster validation techniques were applied to select the most useful model of each type, which were then combined to select candidate records for physical examination by DFAS auditors. The unsupervised modeling techniques utilized available valid transaction data, much of which is not admitted under the current supervised modeling procedure. He demonstrated a new clustering approach called Tree Clustering, which used Classification

and Regression Trees to cluster data with automatic variable selection and scaling.

A Knowledge Base of Erroneous Payment data has been collected at DFAS IR and it is this data base that will be evaluated to further enhance the auditing process at DFAS IR. It is important to develop sound auditing techniques both in the private and public sectors to achieve cost savings and to minimize fraudulent behavior.

# III. BACKGROUND

## A. BACKGROUND INFORMATION

### 1. Genesis of DFAS IR Seaside

DFAS is one of the largest accounting agencies in the world, disbursing nearly one billion dollars every business day. DFAS was formed in January 1991 to eliminate redundant disbursement activities within the Defense Department. In Jenkins [2], Evaluation of Fraud Detection Data Mining Used in the Auditing Process of the Defense Finance and Accounting Service, it was noted that prior to DFAS's inception, DOD had 338 accounting and finance offices worldwide. This excessive number of systems and personnel cost the government 3.1 billion dollars per year in fixed overhead. In addition to this overhead, the large bureaucracy and the lack of standardization left the Defense Department vulnerable to fraud.

Most of the fraud cases found during the early and mid-1990's were discovered by accident, a situation that pointed to systematic problems in the DOD payment system. [1] This problem has been continually addressed since that time with improved internal controls, operational audits and system standardization. However, more proactive techniques were needed to actively fight fraudulent activity. In 1994 Congress created a new unit, called Operation Mongoose, whose sole purpose was to develop methods to detect and prevent fraud. [1] After some reorganization, in the late 1990's, Operation Mongoose became the Seaside branch of DFAS IR. DFAS IR agents work closely with the Defense Manpower and Data Center (DMDC)

agency to gather data appropriate for analysis. DFAS IR assists the audit process with data analysis by searching for problem transactions such as duplicate payments, overpayments and fraud. The synergy developed by tying together these multi-agency functions has resulted in millions of dollars in duplicate payments being recovered, the initiation of fraudulent payment investigations and the improved ability of auditors to identify Conditions Needing Improvement at DFAS payment centers. [3]

## B. DESCRIPTION OF DFAS IR SEASIDE DATA MINING

### 1. Overview of IR Seaside's Analytical Procedures

IR Seaside uses several different analytical techniques to identify problem payments. Before each site audit is conducted, the preceding eighteen months of site data is compiled for analysis. The IR audit coordinator will decide how many records will be selected for screening for a detailed audit with a typical breakdown of 30% duplicate payments, 30% supervised records, 10% unsupervised records, 20% related records and 10% random records. A brief description of each technique used is given below:

#### *a. Supervised Modeling*

The data miners use a knowledge base of fraudulent and erroneous payments to build predictive models. [2] Using this information they are able to develop models to aid in predicting erroneous payments that could lead to detecting fraudulent behaviors.

#### *b. Unsupervised Modeling*

This type of modeling covers all other areas not covered by supervised modeling. Some of the techniques used to date include clustering and pseudo-supervised clustering.

### *c. Duplicate Payments*

These types of payments are made to a vendor, under a valid contract, that has already been paid. DFAS IR currently has five fraud payment models that are used for selection of records to be audited. The models are set up such that if two or more records have the same payment field information, then there is a good chance that the payments were erroneous, fraudulent or both.

(1) DM0102 – Duplicate Payments. This model compares the Purchase Item Identification Number, Delivery Order Number, Invoice Number, Invoice Amount and Disbursing Office Voucher Amount plus Discount minus Interest payment fields.

(2) DM0109 – Duplicate Payments. This model compares the Purchase Item Identification Number, Delivery Order Number, Invoice Number, Invoice Amount, Disbursing Office Voucher Amount plus Discount minus Interest and Merchandise Delivery Date payment fields. The difference from the DM0102 model is that the Merchandise Delivery Date is compared, also.

(3) DM0110 – Exigency Contract Duplicate Payments. This model compares the Invoice Number, Invoice Amount and Disbursing Office Voucher Amount plus Discount minus Interest payment fields. The difference from the DM0102 model is that the Purchase Item Identification Number and Delivery Order Number are not compared.

(4) DM0111 – High Dollar Duplicate Invoice Amount. This model looks at Purchase Item Identification Number, Delivery Order Number, Same Invoice Number, Invoice Amount and Disbursing Office Voucher Amount plus Discount minus Interest. The difference from the DM0102 model is that the Invoice Number is not compared.

(5) DM0210 – Exigency Contract Duplicate Payments. This model compares the Invoice Amount, Invoice Date, Invoice Amount (greater than $200.00) and Disbursing Office Voucher Amount plus Discount minus Interest payment fields. The difference from the DM0102 model is that the Purchase Item Identification Number, Invoice Number and Delivery Order Number are not compared. The DM0102 Duplicate Payment Model will be analyzed in this thesis.

To identify duplicate payments the DFAS IR team evaluates all payments made at a DFAS site. This technique is initially computer-intensive in the comparison of all records with specific matching rules developed by the IR auditors. Site records are compared pair-wise and several new record fields are generated. The new fields indicate whether a record shares traits in common with another record in the database. If two records are nearly identical then they are flagged as a potential erroneous payment. The DFAS IR auditors apply the five fraud models to determine whether records deserve attention during upcoming site visits. Duplicate payments have been the most productive and visible aspect of their data mining work with over $75 million dollars recovered to date. [3]

### *d. Related Payments*

These payments are records that are "related" to the records selected by the supervised models. When the supervised modeling process selects a record, a Defense Manpower Data Center (DMDC) query then finds all other records related to the suspect record in the fields of payee, contract, address, or electronic fund transfer number. All the related records are documented and the information is brought to the site audit. The related

records may or may not be reviewed during the site audit depending on whether the audit reveals problems with its associated supervised record or if the data mining team deems the record to be interesting. [2]

#### *e. Random Records*

This method has been the traditional way of choosing records for DFAS audits. DFAS IR accomplishes random selection by assigning each record a random number from one to the number of records. The records are sorted by random numbers and the records with smallest numbers are selected until the desired number of records is obtained.

### 2. In-Depth Review of Supervised Modeling

Figure 1 shows the logical process for choosing the supervised versus unsupervised, random and duplicate record selection. This is how DFAS IR prepares for each pay site audit. The remainder of this section will explain the major steps of the Record Selection Process in more detail.

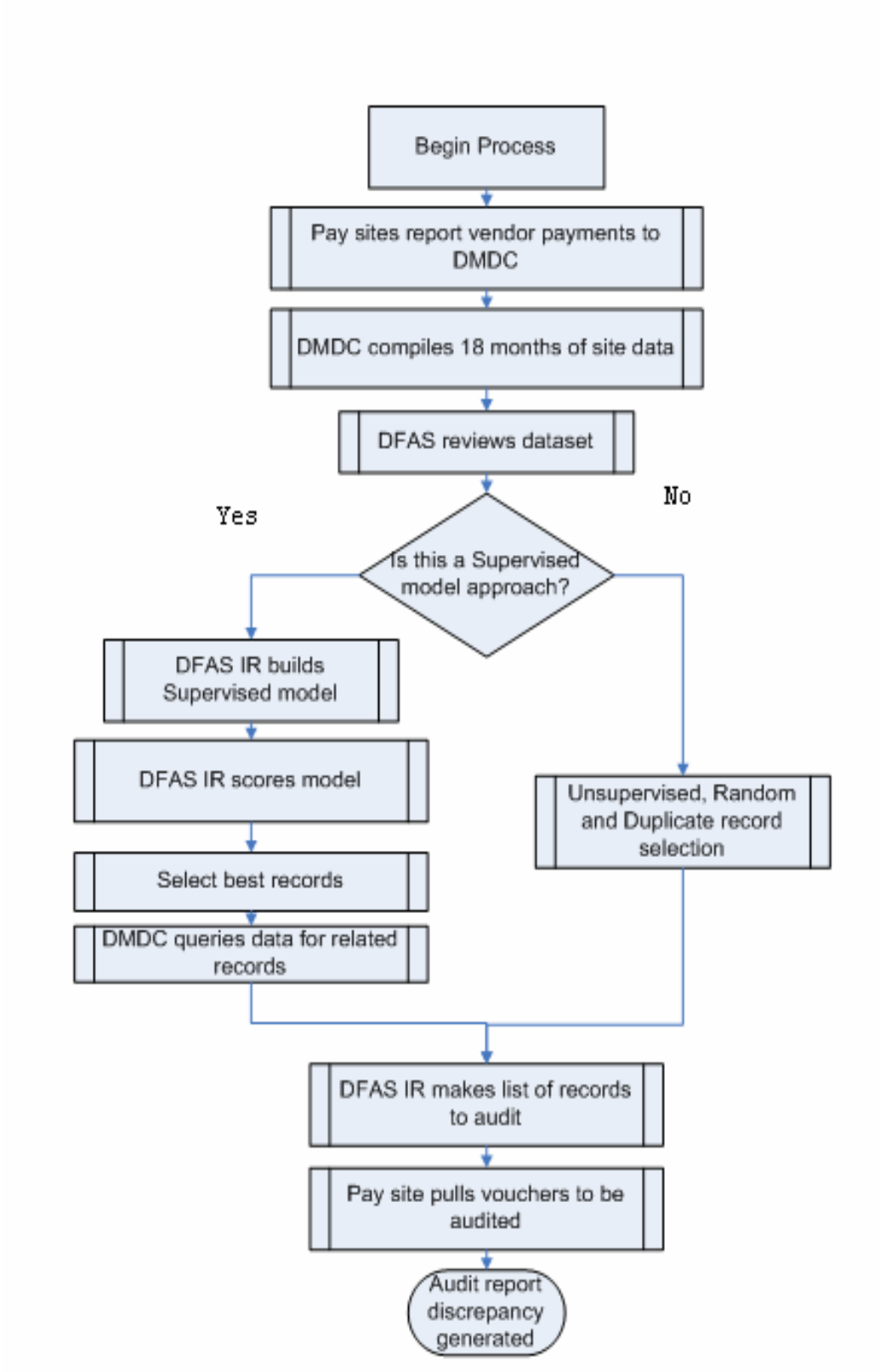



Figure 1. Record Selection Process Flowchart

### *a. Description of Fraud Knowledge Base*

The fraud knowledge base used in the supervised model comes from the known prosecuted fraudulent activity. This knowledge base consists of 453 transactions from 16 known fraud cases. [6] At the beginning of Operation Mongoose, these records from cases were collected from DMDC or from the actual transactions. The data was analyzed using principal component analysis along with clustering techniques to group the payments for easier classification. [6] This resulted in the fraudulent payments being broken down into four fraud "types" for modeling purposes. These types were labeled as *Big Systematic (BigSys), Small Systematic (SmallSys), Opportunistic (Ops), and Piggyback (Piggy)*.[2] The difference between *BigSys* and *SmallSys* is the dollar amount, but both are due to errors in the procedural payments on behalf of DFAS pay sites to the vendors. *Opportunistic payments* are where the vendor resubmits for a manual payment after having already received an electronic payment. The *Piggyback* scenario occurs when the vendor will try to submit for a consolidated payment under one request and then request a payment on one of the smaller line items within a contract with a separate request. An example of this is when there is an extra charge for exceeding a household goods weight requirement. The charge is lumped into the initial request and then a separate invoice is made for the amount that exceeds the weight requirement.

### *b. Site Data Review and Preparation*

As with any application, it is important to verify data integrity before performing an analysis. Prior to modeling any site data, the statistical information for each field is compiled and reviewed by the senior data-

miner. The senior data-miner then releases a comprehensive spreadsheet with recommendations as to which fields to avoid or to use in model development. Serious data integrity issues are reported to DMDC so the responsible site can be informed of the data entry problems that need to be addressed at the site. [2] This initial step of cleaning prevents the analysis from having to be repeated later. The next step is to break the data into three subsets: training, test and validation. The training set is used to construct models, which are then tested using the test set and evaluated using the validation set. [2]

### *c. Model Building and Scoring Process*

The data is divided and distributed to several data miners to disperse the workload for model development. The data miners use the Clementine data mining software to build the models using methods ranging from linear and logistic regression to neural networks. The data mining efforts have primarily used classification trees for their ease of understanding and neural networks for their ease of use. [7] At this point the data miners will use their knowledge of previous site audits to build what they consider the best models and provide a scoring for each of the models that are developed.

When the data mining project began, the analysts lacked feedback regarding the effectiveness of their models. Therefore, they designed an *ad hoc* scoring function that has been in use ever since. After building a satisfactory model on the training data, the modelers run the test and validation datasets through the model. The results are contingency tables with counts distributed by rows of known fraud status and columns of predicted fraud

status for the test and validation sets. One major assumption is made when applying the known fraud label: that none of the sampled site data is fraudulent. Given the large number of payments and the belief that most payments are not fraudulent, this assumption is reasonable. [2] The data miners are able to compare all models on one single spreadsheet. The scores are then used as an objective factor in the subjective selection of models for the site's supervised ensemble.

### *d. Model Ensemble Collection*

All the models are gathered together and selected based on the objective score, along with the intent to evenly distribute splits, modelers and classification methods. After the models are selected this information is put into Clementine and the software selects the records to be chosen for auditing.

### *e. Record Selection for Audits*

Once the modeling process has been completed, the entire 18-month database of site records is run through the model ensemble. Each model classifies each record and the predicted fraud classifications for each record are counted. A true simple majority-voting scheme would classify all records that receive a majority vote as potentially fraudulent and worth review. However, audit team resources and time are limited, so only a fixed number of records can be selected. [2]

### *f. Audit Preparation*

Records selected by the different techniques are referred to as candidates. The data mining team sends the candidate list back to DMDC. DMDC then prepares and returns a list of the candidates and any related records to DFAS IR. Approximately two weeks prior to the site visit,

DFAS IR forwards the candidate list to the audit site so that the record's documentation can be prepared for presentation upon the audit team's arrival. [2]

### 3. In-Depth Review of Unsupervised Modeling

An important class of unsupervised learning is cluster analysis or data segmentation. Cluster analysis is used to describe a data set in its entirety, grouping together similar observations into distinct clusters. The "distance" between clusters depends on their degree of dissimilarity; observations that fall into two clusters that are "close together" are more similar to one another than observations from clusters that are "far apart." Some measure of the similarity between observations must be calculated in order to find clusters in the data set. Most clustering algorithms utilize a numeric matrix (called a similarity or dissimilarity matrix) to represent the distances between observations. Thus any non-numeric variables must be coded numerically in terms of similarity or dissimilarity. The reader interested in a detailed account of unsupervised modeling is referred to Hand [18].

#### *a. Shortcomings of Supervised Modeling*

The primary shortcoming of the supervised modeling methodology currently in place is that models are developed from possibly outdated, incomplete or potentially misclassified Knowledge Base (KB) information. Additionally, the supervised modelers at Operation Mongoose work very hard to create complex models and combinations of models that consistently perfectly predict all the KB transactions of a particular type. Success with the KB transactions is unlikely to translate into success for new transactions for such complex models.

Although the population data is randomly divided among the data splits, the assignment of KB transactions to the data splits is predetermined. This brings into question the validity of the predictions made by the resulting models. [8] Given historical trends of erroneous and fraudulent behavior, the modelers are trying to focus their analysis on specific payment fields. This approach may overlook other relationships that may be useful in predicting fraudulent or erroneous payments.

### *b. Potential Improvements with Unsupervised Modeling*

The primary potential improvement with unsupervised modeling is the ability to exploit all the data in the population without regard to the Knowledge Base. Additionally, an unsupervised model may reveal actual patterns in the population data, independent of the preconceived (and potentially incorrect) fraud classifications in the KB. [8]

## C. ANALYSIS OF PREVIOUS SITE AUDITS

This analysis considers five pay activities in the DFAS payment system; the findings are displayed in Table 1. The total duplicate payment shows the result of each payment sites audit. The percentage of transactions that were found to have erroneous payments is relatively small ranging from .01 % to 1.32 %.

| DFAS Payment Site | Period of Review | Total disbursed (Billions) | | Total Duplicate Payments Identified | Value | Percentage of Disbursed |
|---|---|---|---|---|---|---|
| DFAS Charleston | Jun-Dec03 | $ | 1.30 | 80 | $6,471,482.00 | 0.50% |
| DFAS Columbus | Apr-Dec 2001 | $ | 1.50 | 60 | $1,297,900.00 | 0.09% |
| DFAS Columbus | Oct01-Sep02 | $ | 83.60 | 70 | $8,370,296.00 | 0.01% |
| DFAS Kansas City | Mar-Jun03 | $ | 0.47 | 96 | $6,058,036.00 | 1.30% |
| DFAS Pacific | Dec02-Mar03 | $ | 3.60 | 85 | $1,588,514.00 | 0.04% |

Table 1. Site Audit Data

### 1. Site Discrepancies and Corrections

#### *a. DFAS Charleston (June-December 03) [11]*

(1) Description of Audit Discrepancies. DFAS Internal Review identified erroneous payments primarily caused by either certification office errors or data input errors. DFAS Charleston's vendor pay management was unable to obtain adequate supporting documents for 80 potentially erroneous payments processed by DFAS Charleston. The structure of duties related to running the duplicate payment query and investigating the possible duplicates was not consistent with the standard structure developed.

(2) Correction. Recovered overpayments and increased vigilance from management.

#### *b. DFAS Columbus (DFAS CO) (April-December 2001) [9]*

(1) Description of Audit Discrepancies. The most common erroneous payment was due to a lack of audit controls for duplicate FEDEX payments which caused the payment office not to be aware of 11 of 60 duplicate payments, totaling $37,643. Offsets were made against pending invoices at Federal Express (FEDEX) and Air Force Materiel Command (AFMC), rather than forwarded to DFAS CO to be entered into the payment system as debits and credits. Therefore, the payment office has no valid audit trail for disbursements and collections that have been offset by FEDEX or AFMC.

(2) Correction. Recovered overpayments and ensured that DFAS CO receives invoices for entering into the payment system and keeping DFAS CO in the payment process.

### *c. DFAS CO (October 01-September 02) [13]*

This audit dealt with identifying problems in a large multi-million dollar pre-payment contracts.

(1) Description of Audit Discrepancies. The cause of most overpayments involved duplicate invoicing by contractors of additional fabricated shipments. The current system's edit and prepayment reports were not designed to detect duplicate payments involving different shipment numbers where the Invoice Number is not the same.

(2) Correction. Internal Review Seaside has developed fraud detection models that can identify these potential overpayments. Like previous models developed with the collaboration of the DFAS CO Systems and the Quality Directorate, the model is intended to assist in the creation of a prepayment report that will help to detect these payments with particular characteristics before payment.

### *d. DFAS Kansas City (March-June 03) [12]*

(1) Description of Audit Discrepancies. The Vendor Pay managers indicated that duplicate payments resulted from voucher examiner errors, duplicate invoice submissions by the activity, certification office error, contracting officer error, data input error, system problems and Electronic Fund Transfer accounts expiring.

(2) Correction. Recovered overpayments and increased vigilance from management in the above-mentioned areas.

### *e. DFAS Pacific (December 02 – March 03) [10]*

(1) Description of Audit Discrepancies. Incorrectly entered dates caused duplicate payment edits to fail and not be properly identified as erroneous payments.

Both vendors and payment activities submitted duplicate invoices, which caused the duplicate errors.

(2) Correction. Recovered overpayments and provide training for personnel to look through dataset of invoices to determine possible duplicate payments.

**2. Final Analysis**

A common theme throughout the five site audits was that the majority of erroneous payments were due to a lack of controls and oversight at the DFAS payment sites. As noted in Table 1, the percentage of erroneous payments captured ranged from .01 to 1.32 % of the total disbursed. This appears to be an insignificant amount in comparison to the total disbursed, but total dollar value of erroneous payments recouped was roughly $ 23.79 million.

Are the techniques being used to determine fraudulent and erroneous payments effective? Although these techniques have recouped these overpayments, the degree to which they are effective cannot be determined. As time passes and data is collected, will the effectiveness of these techniques be determined? Vigilance and sound metrics by management and DFAS IR will assist in these efforts. Most of the problems with the erroneous payments seem to be a function of poor audit and quality assurance checks by management at the DFAS pay sites. Each activity can look at placing more emphasis on detecting fraudulent vendor payment patterns by applying rigorous training and auditing of erroneous payment transactions.

# IV. RESEARCH METHODOLOGY

## A. ANALYSIS OVERVIEW

DFAS IR auditors now have a Knowledge Base of Erroneous Payments (KBOEP) have been verified by actual on-site audits. This database identifies those payments that money was recouped as a result of these audits. Data mining at DFAS IR has progressed to the point where the staff is efficient at data review, modeling and record selection. [2] The goal is to reduce the time required to manually select records by developing software models to pre-select records for site audits. Fraud prosecution is very important, but it will not be the main motivation behind developing these models. Jenkins [2] states, "Because of the long time required between identifying potentially fraudulent records, investigation and prosecution, it is impractical to use this as a performance measure." Davia [9] resoundingly rejects fraud detection as a performance measure because of the historical difficulty of prosecuting fraud. He points out that proactive fraud auditing's greatest strength lies not in its ability to detect fraud, but more in its deterrent aspects. If the selection processing of very large data set can be improved and the manual time spent looking through records reduced, then the auditors may spend more time focusing on looking at fraudulent, behavior patterns.

The next step for DFAS IR is to take this KBOEP and determine if they can replace their current manual pre-audit record selection with an electronic record selection process. Several statistical techniques will be used to look for trends within the KBOEP. The current database

consists of records that have been selected and audited, and money recovered, from various payment sites. The records have been labeled with a 1 (Success) when the record produced an erroneous payment and a 0 (Failure) for those that did not. This field name is called the 'Target' and will be referenced throughout. Logistic Regression (supervised modeling) along with the Hosmer-Lemeshow Test for validity, Classification and Regression Trees (C&RT) and Receiver Operating Characteristic (ROC) curves are used to develop and test the strength of the predictive models. Appendix A shows the fields to be used in the analysis.

## B. TOOLS USED FOR ANALYSIS

### 1. Logistic Regression

#### *a. Overview*

Many questions in science involve trying to predict the probability that something will happen, for example, the probability that people will vote for one of two candidates or that someone will have AIDS. Such questions involve two-category (dichotomous) *Y* variables for example, vote/no vote or AIDS/no AIDS. In this thesis the dichotomous relationship is erroneous payment/non-erroneous payment. The erroneous payment will have a *Y* value of 1 or 0.

#### *b. Logistic Regression Construction*

Hamilton [15] shows that the simple linear regression model is appropriate for relating a quantitative response variable *Y* to a quantitative predictor *X*. Suppose that *Y* is a dichotomous variable with possible values 1 and 0 corresponding to Success and Failure, and let P=*P*(*Y*=1). The value of P will depend on the value of some quantitative variable *X*. For example, the probability that a car needs warranty service of a certain kind might well

depend on the car's mileage or the probability of avoiding an infection of a certain type might depend on the dosage in an inoculation. Instead of using just the symbol *p* for the success probability, now use *p*(*X*) to emphasize the dependence of this probability on the value of *X*. The simple linear regression equation

$$\hat{Y} = \beta_o + \beta_1 X + \varepsilon \qquad (4.1)$$

is no longer appropriate.

P(*Y*=1) denotes the probability that a {0,1} *Y* variable equals 1. The probability that *Y* does not equal 1 is

$$P(Y \neq 1) = P(Y = 0) = 1 - P(Y = 1) \qquad (4.2)$$

The odds favoring *Y* = 1 are

$$\Theta(Y = 1) = \frac{P(Y = 1)}{1 - P(Y = 1)} \qquad (4.3)$$

Odds range from 0(when P(*Y*=1)=0 to ∞ (when P(*Y*=1)=1).

Suppose Y = 1 indicates that it rains today and Y = 0 indicates that it does not. If the probability of rain today is P(Y=1) = 0.2, then the probability of no rain is 1-P(Y=1)=0.8. The odds of rain today are

$$= \frac{0.2}{0.8} = \frac{1}{4} = 0.25 \qquad (4.4)$$

These odds could be stated as 0.25 to 1 or 1 to 4. Thus a 0.25 probability amounts to 1-to-4 odds. By taking the natural logarithm of the odds, we obtain a logit:

$$L = \log_e \Theta = \log_e \left\{ \frac{(P)}{(1-P)} \right\} \qquad (4.5)$$

Logits range from -∞ (when P=0) to ∞ (when P=1). Logit regression refers to models with a logit as left-hand-side variable:

$$L_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{K-1} X_{i2,K-1} \qquad (4.6)$$

If the logit (L) is a linear function of X variables, then the probability (P) is a non-linear, S-shaped function like that in Figure 2. Predicted probabilities approach, but never reach or exceed, the boundaries of 0 and 1. Thus logit regression provides a more realistic model for probabilities than linear regression does.

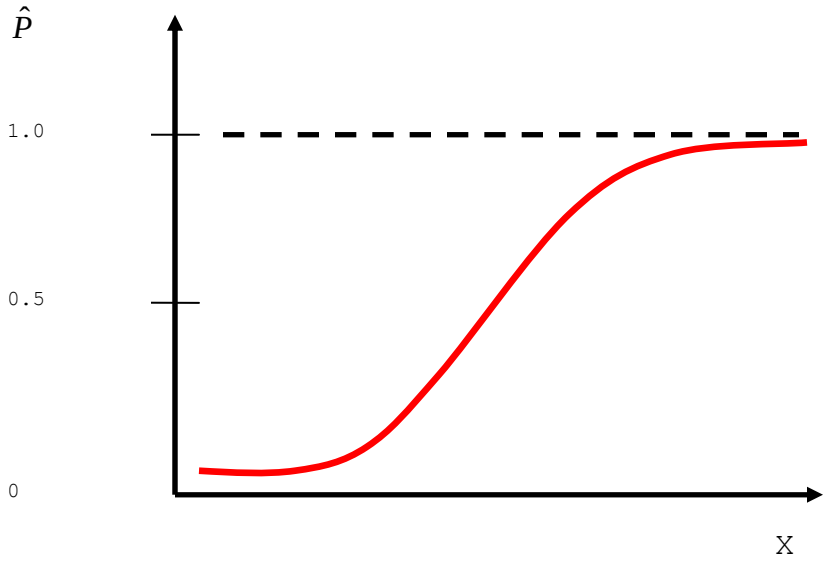



Figure 2. A Graph of a Logit Function

Given a set of X values and estimated coefficients, we can estimate logits ($\hat{L}$) much as we do $\hat{Y}$ in linear regression. Reversing the logit transformation yields predicted probabilities that Y=1($\hat{P}$):

$$\hat{P} = \frac{1}{1+e^{-\hat{L}}} \tag{4.7}$$

This re-expression is useful for graphing. Unless X strongly affects Y, graphing $\hat{P}$ over the data's X range will not show a complete S-curve; instead it will be a partial curve of Figure 2.

#### *c. Vendor Payment Knowledge Base Application*

A Logistic Regression model will be developed by separating the Fraud Type Indicator DM0102 from the KBOEP data into a training and test set. The training set will generate a model for the separated data and will be compared against the test set for accuracy of predictions.

### 2. Hosmer-Lemeshow Test

#### *a. Overview*

Hosmer and Lemeshow [20] provide an interesting approach to evaluating the quality of logistic regression. After a model is built a predicted probability is generated for every observation. These predictions are sorted and divided into *g* groups of approximately equal size according to their predicted probability. Once the groupings are done a Chi-square goodness-of-fit test is performed. Using an extensive set of simulations, they demonstrated that when the groupings are done properly and the fitted logistic regression model is the correct model, the test statistic is well approximated by the Chi-square distribution with *g*-2 degrees of freedom, $\chi^2_{(g-2)}$. [20]

### *b. Hosmer-Lemeshow Construction*

Separating the estimated probabilities into n columns, where the first column corresponds to the smallest value and the nth column to the largest value. Two grouping strategies were proposed: (1) collapse the table based on percentiles of the estimated probabilities and (2) collapse the table based on fixed values of the estimated probability. [20]

With the first method, use of $g$ = 10 groups results in the first group containing the $n_1' = n/10$ subjects having the smallest estimated probabilities and the last group containing the $n_{10}' = n/10$ subjects having the largest estimated probabilities. With the second method, use of $g$ = 10 groups results in cutpoints defined at the values $k/10$, $k$ = 1, 2… 9 and the groups contain all subjects with the estimated probabilities between adjacent cutpoints. For example, the first group contains all subjects whose estimated probability is less than or equal to 0.1, while the tenth group contains those subjects whose estimated probability is greater than 0.9. Now, building a new table with two rows, with a $y$ = 1 row, estimating the expected value obtained by summing the estimated probabilities over all subjects in a group and a $y$ = 0 row, estimating the expected value of obtained by summing, over all subjects in the group, one minus the estimated probability. For either grouping strategy, the Hosmer-Lemeshow goodness-of-fit statistic, $\hat{C}$, is obtained by calculating the Pearson Chi-square statistic from the 2 x $g$ table of observed and estimated expected frequencies. A formula defining the calculation of $\hat{C}$ is as follows:

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n_k' \bar{\pi}_k)^2}{n_k' \bar{\pi}_k (1 - \bar{\pi}_k)} \qquad (4.8)$$

where $n_k$' is the number of covariate patterns in the $k^{th}$ group,

$$o_k = \sum_{j=1}^{n_k'} y_j \qquad (4.9)$$

is the number of successes among the $n_k$' covariate patterns and

$$\bar{\pi}_k = \sum_{j=1}^{n_k'} \frac{m_j \hat{\pi}_j}{n_k'} \qquad (4.10)$$

is the average estimated probability within group k. [20]

At this point we have a table of observed and estimated expected values within each grouping. The Hosmer-Lemeshow goodness-of-fit test statistic is computed from the frequencies in the table. A $\chi^2$ *p*-value close to one is indicative of a good-fitting model. The best $\chi^2$ for a given dataset may be more a function of the environment being tested in some cases may not be close to 1.0. A more detailed development of the Hosmer-Lemeshow and the $\chi^2$ goodness-of-fit test can be found in [20].

### *c. Vendor Payment Knowledge Base Application*

Once the Logistic Model has generated the predicted probabilities, the Hosmer-Lemeshow Test, using a function built by Professor Samuel Buttrey, is run to determine the goodness-of-fit. See Appendix B for Clementine nodes and stream used.

### 3. Classification Trees

#### *a. Overview*

As described in [18], The Classification and Regression Trees (C&RT) algorithm is a widely used statistical procedure for producing Classification and Regression models with a tree-based structure. They are non-parametric supervised procedures to explain and predict the response variable based on one or more input variables. For this discussion, consider only the classification aspect of C&RT, which is mapping an input vector *X* to a categorical (class) output label *Y*.

#### *b. C&RT Construction*

The structure of the tree is derived from the data; C&RT works by choosing the best variable for splitting the data into two groups at the root node. It can use any of several different splitting criteria; all produce the effect of partitioning the data at an internal node into two disjoint subsets (branches) in such a way that the class labels in each subset are as homogeneous as possible This splitting procedure is then recursively applied to the data in each of the child nodes and so on. The size of the final tree is a result of a relatively complicated "pruning", process, outlined in [18], chapter 5.

#### *c. Vendor Payment Knowledge Base Application*

The type of C&RT that will be used in the Clementine software will be the C5.0 algorithm. It builds a decision tree or ruleset, by splitting the sample based on the field that provides the maximum information gain. Each subsample defined by the first split is then split again, usually based on a different field, and the process repeats until the subsamples cannot be split any further. Finally,

the lowest level splits are reexamined and those that do not contribute significantly to the value of the model are removed or pruned.

The uniqueness of the C5.0 algorithm is that it can provide two kinds of models, a decision tree and a ruleset. The decision tree is straightforward, because it provides a description of the data by separating the data into respective terminal, or "leaf" nodes, each describing a particular subset of training data. Any observation in the training data belongs to exactly one terminal node in the tree. The Target variable is the detection of an actual erroneous payment found by auditors at DFAS payment sites, a one if an erroneous payment is found and a zero otherwise. In Figure 3, the node graph shows a portion of a generated symbolic target field. The graph is a chart of percentages in each category of the Target field. Preceding each row in the table is a color swatch that corresponds to the color that represents each of the target field categories in the graphs for the node. In this case a zero (light blue) or one (red).

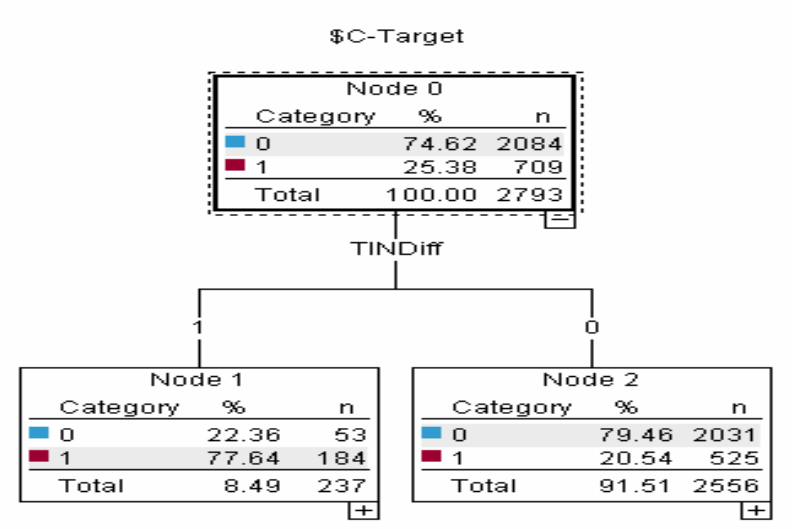



Figure 3. Classification and Regression Tree diagram

The Taxpayer Identification Number Difference field (TINDiff) has been selected as the first field to begin with. The chart within the Node 0 box indicates the proportion of zeroes and ones and their relationship to Target variable. Node 1 represents the Invoice Date Difference (InvDtDiff) field's relationship to TINDIFF and the Target variable. Node 2 represents the Check Date Difference (ChkDtDiff) field's relationship to TINDiff and the Target variable. The splits can be divided into a minimum of two to a maximum of eight subsets in the Clementine software.

Clementine can also generate a ruleset for use in selecting records can be selected electronically. The Clementine User's Guide [7] defines a ruleset as a set of rules that tries to make predictions for individual records. Rulesets are derived from decision trees and, in a way, represent a simplified or distilled version for the information found in the decision tree. Rulesets can often retain most of the important information from a full decision tree but with a less complex model. Rulesets do not have all of the same properties as decision trees. The most important difference is that with a ruleset, more than one rule may apply for any particular record, or no rules at all may apply. If multiple rules apply, each rule gets a weighted "vote" based on the confidence associated with that rule and the final prediction is decided by combining the weighted votes of all of the rules that apply to the record in question. If no rule applies, a default prediction is assigned to the record. The Clementine Users Guide [7] provides a detailed explanation of selection options shown in Figure 4.

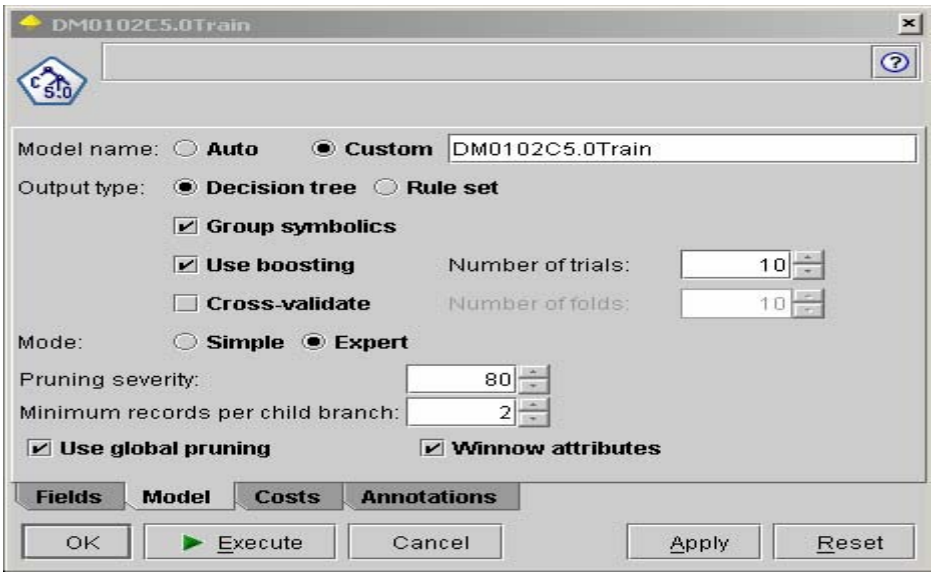



Figure 4. Setting Options for an Evaluation Chart Node

#### *d. Vendor Payment Knowledge Base Application*

A C&RT model was developed by separating the DM0102 KBOEP data into a training and test data set. The training set was set aside in order to develop a model to predict the 'Target' variable. The pruning severity and minimum records per child branch were adjusted to see which mix will provide the best result. This is akin to setting the threshold level for splitting the tree into its respective branches. The test set model was run to compare its predictive ability against the training set.

### 4. Receiver Operating Characteristic Curves (ROC)

#### *a. Overview*

ROC curves were developed in the 1950's as a by-product of research into making sense of radio signals contaminated by noise. More recently they have found usefulness in many statistical applications. [19] The curves are generated to assist in the understanding of the coincidence (confusion) matrices. A coincidence matrix looks like the one in Table 2. In this table, True

Positive Fractions (TPF) represents those records that are erroneous payments and have a "high" test (above whatever cutoff level was chosen). FPF represents false positives, where the test has told us that a non-erroneous payment was really erroneous. The True Negative Fraction (TNF) represents correctly identified Non-Erroneous Payments and the False Negative Fraction (FNF) represents erroneous payments incorrectly classified as being non-erroneous.

| Actual Target vs test | | |
|---|---|---|
| | Erroneous Payment | No Erroneous Payment |
| "high" test (positive) | TPF | FPF |
| "low" test (negative) | FNF | TNF |
| "high" and "low" test refers to value relative to some arbitrary cutoff point. FNF+TPF =1 / TNF + FPF =1 | | |

Table 2. Coincidence Matrix as Fractions

Some coincidence matrices display not the fractions but the actual number of records that were identified. The Clementine software generates coincidence matrices in this manner and the fractions are easily calculated and displayed separately.

### *b. ROC Curve Construction*

Central to the idea of ROC curves is this idea of a cutoff level. A test is declared "positive" if the value is above some arbitrary cutoff, and "negative" if below. An example is shown in Figure 5. The elliptical shape shows the location of the vertical line that intersects the bell shaped curves and this serves as the cutoff point. In Figure 5 the vertical line threshold (Test value>) is very high (to the right of the two bell-shaped curves), which results in almost no *false* positives, and very few *true*

positives, as noted by the circle's position on the ROC curve in Figure 5. Both TPF and FPF will be close to zero, so we are at a point close to (0,0) on the ROC curve.

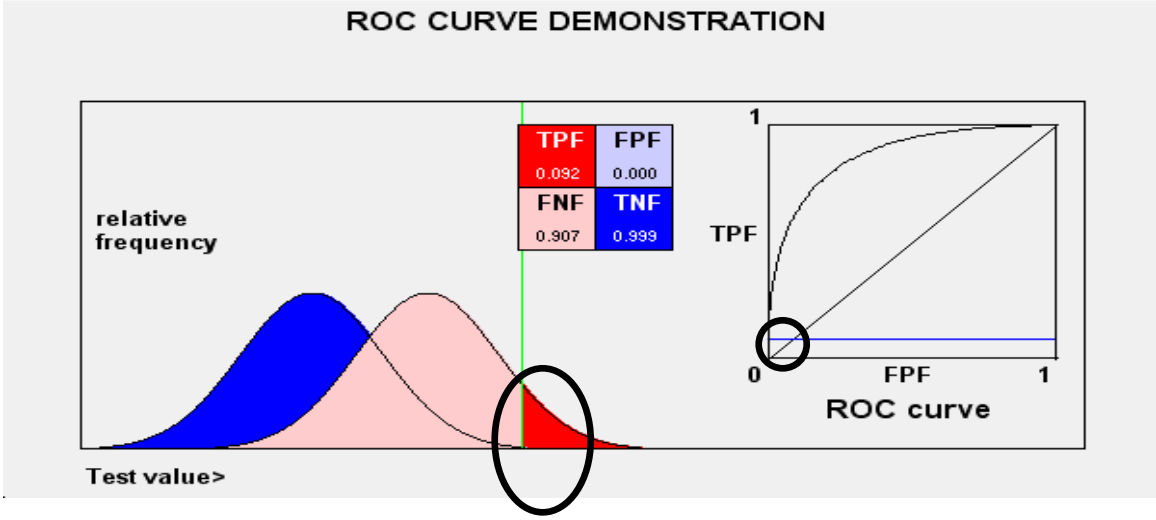



Figure 5. ROC Curve Low Threshold [19]

In Figure 6, we move the vertical line test threshold towards a more reasonable, lower value (to the left), so that the number of true positives will increase (rather dramatically at first, as the ROC curve moves up steeply). Finally, a point is reached on the ROC curve where there is a remarkable increase in *false positives.* The ROC curve slopes off as we move our test threshold down to very low values. Again the vertical line threshold (cutoff) corresponds to ellipse's location on the bell-shaped curves and the circle on the ROC curve.

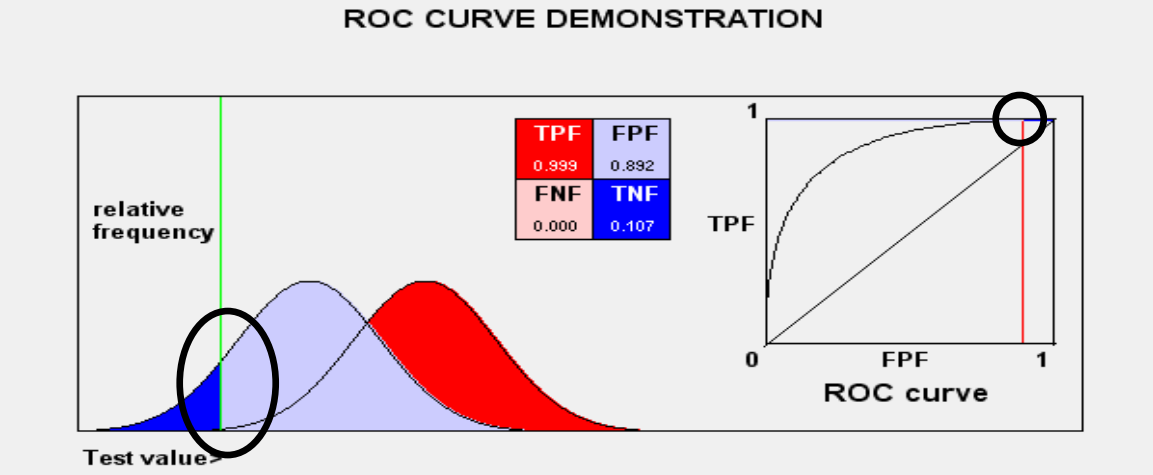



Figure 6. ROC Curve High Threshold [19]

This shows the impact of changing the threshold level. Making the cutoff too high corresponds to lower TPF and FPF, but making it too low results in high TPF and high FPF, both of which are undesirable. Any choice of cutoff level produces a tradeoff in assessing your models ability to predict outcomes. Ideally you want to choose a cutoff point that will give you the highest TPF with the smallest FPF.

The next step is to look at the effect of changing the overlapping bell-shaped curves and its affect on the ROC curve. An example from [19] provides an intuitive feel for the effects the bell-shaped curves have upon the ROC curve. Consider two tests. The first test is *good* at discriminating between patients with and without a disease. This will be test A. The second test is bad at discriminating between patients with or without a disease. This will be test B. Let's examine each:

Test A, Figure 7, shows the bell-shaped curves are now moved apart and the arrows point to the area defined by TPF and FPF, respectively. With this amount of

separation between the bell-shaped curves the ROC curve has a large area between the straight line and the knee shaped curve. This separation provides a higher TPF (0.978) to FPF (0.225) mix. We want both a high TPF and a low FPF, so that the model predicts relatively well. We want to choose a cutoff level on the ROC curve that keeps the TPF high, while keeping the FPF low. This point would be at the knee of the upper curve of the ROC curve. These are the characteristics of a good ROC curve.

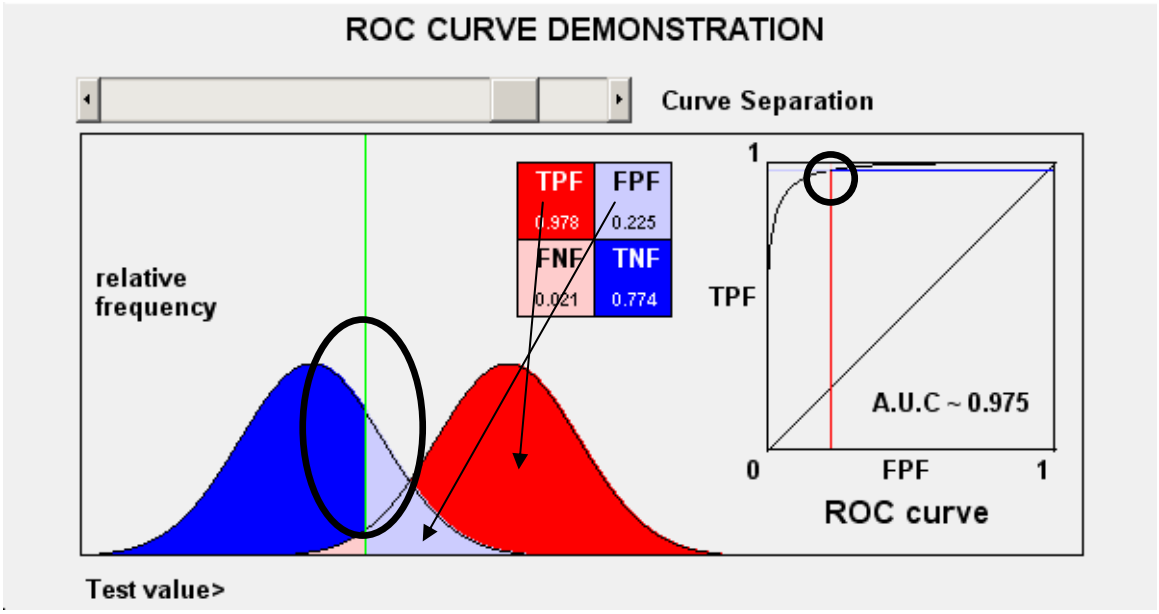



Figure 7. ROC Curve Large Separation [19]

Test B, Figure 8, shows that the bell-shaped curves are almost overlapping and the arrows point to the area defined by TPF and FPF, respectively. Because of the lack of separation of the bell-shaped curves, the ROC curves are relatively close to one another. As we plot the graph on the ROC curve we can see that for every true positive that moves up we are likely to encounter a false positive that moves us to the right. This results in more or less of a diagonal line from the bottom left corner of

the ROC curve, up to the top right corner. The model's ability to discern between a TPF or FPF will almost be equal. These are the characteristics of a poor test.

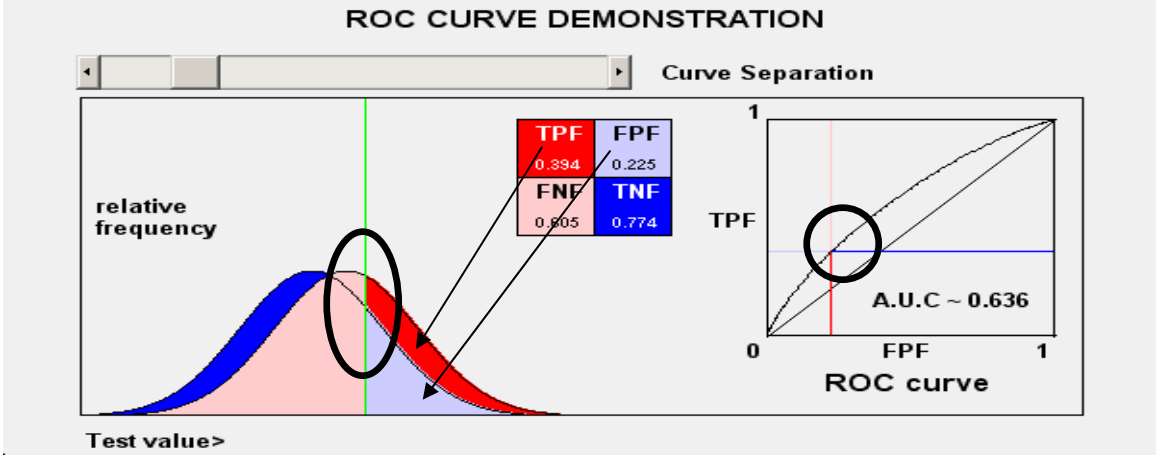



Figure 8. ROC Curve Small Separation [19]

From the above scenario you can get a good intuitive feel that the closer the ROC curve is to a diagonal, the less useful the test is at discriminating between the two populations. The more steeply the curve moves up and then (only later) across, the better the test. A more precise way of characterizing this "closeness to the diagonal" is simply to look at the area under the ROC curve. The closer the area is to 0.5, the poorer the test performs, and the closer it is to 1.0, the better the test performs.

### *c. Vendor Payment Knowledge Base Application*

In the Clementine User's Guide [7], the ROC curve is generated in the Analysis Node and is called the Gain curve. See Figure 9, for the Evaluation Node and the node options display.

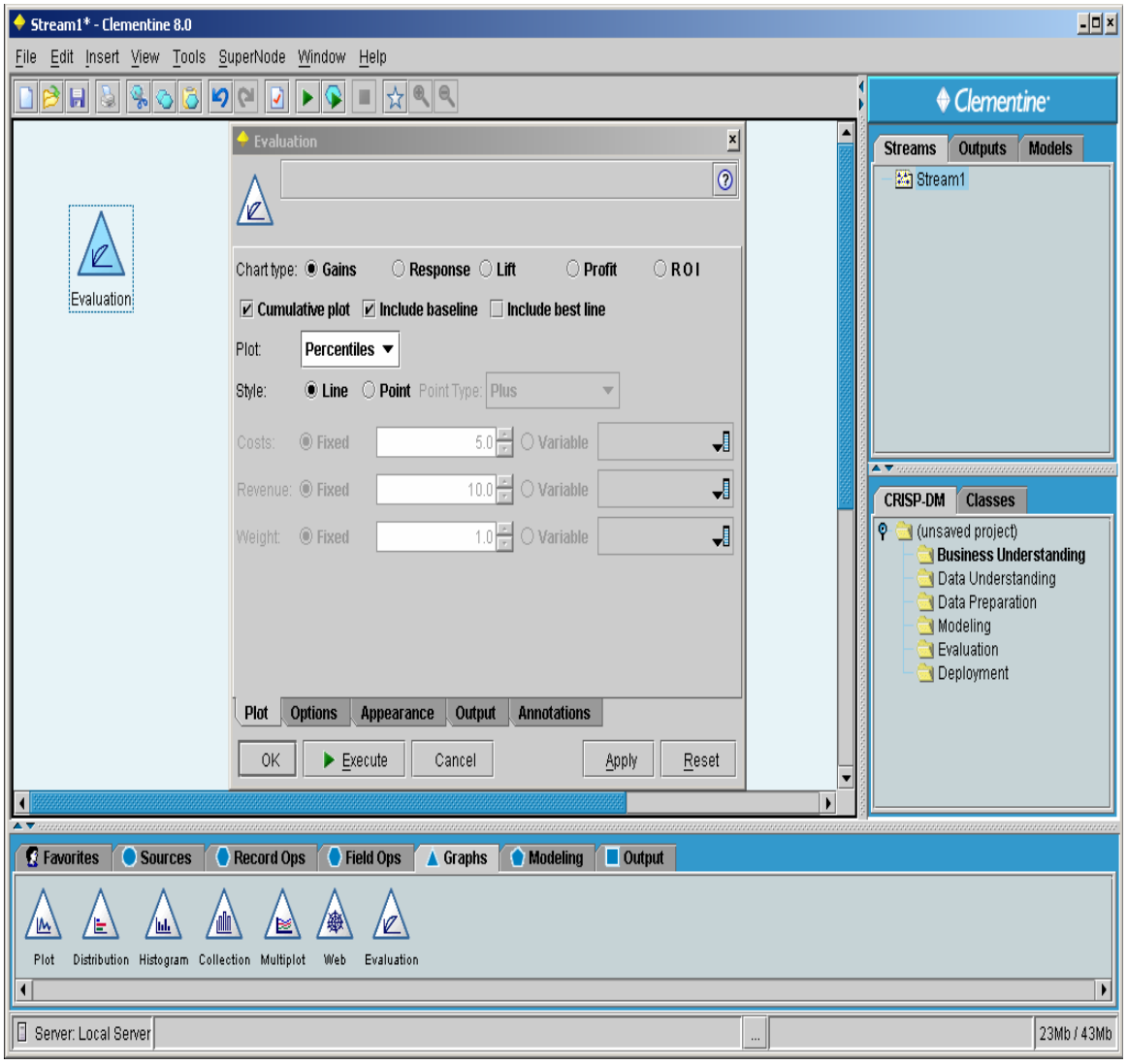



Figure 9. Clementine Evaluation Node

### *d. Implementation*

We run the Logistic Regression and C&RT nodes to develop a model to generate the probabilities for predicting the Target variable of 1 or 0. Once the probability results are generated the Evaluation Node is run and an ROC curve is developed for both analysis tools. Chapter V examines the performance of the two models.

THIS PAGE LEFT INTENTIONALLY BLANK

# V. ANALYSIS

## A. OVERVIEW

The analysis that follows will consist of a discussion of the Logistic Regression, Hosmer-Lemeshow Test performed in the statistical software package S-Plus and Logistic Regression and C&RT analysis in the Clementine software package.

The section on Logistic Regression analysis in the S-Plus software will discuss the model generated and the results from the Hosmer-Lemeshow Test. The section on analysis in the Clementine software will begin with Logistic Regression and C&RT applied to the DM0102 dataset. The resultant analysis will compare each model's ability to predict the Target field's binary outcome. In addition to this analysis, a comparison of the Receiver Operator Characteristic curves will be done to assist in providing insight into the models ability to generate the TPF, TNF, FPF and FNF.

Once a statistical tool is selected, the data set will be further examined to see how it performs individually by breaking the DM0102 dataset into Training and Test set.

## B. LOGISTIC REGRESSION HOSMER-LEMESHOW TEST

A Logistic Regression model was developed in the statistical software package S-Plus and this model was tested using the Hosmer-Lemeshow Test. The test showed that the $\chi^2$ goodness-of-fit was zero. This was an indication that the model that was generated would not perform well. Some attempts were made to eliminate fields and models were generated again and tested with the same resultant zero goodness-of-fit. We conclude that the logistic regression

models were not predicting probabilities of duplicate payment accurately. The Hosmer-Lemeshow Test does seem, at least roughly, to rank the predictions properly. This information would be useful in assessing thresholds of the model's ability to predict. This test was abandoned and other statistical tools were pursued to aid in determining the model's ability to predict.

## C. LOGISTIC REGRESSION

In building this model the dependent listed variable is the *Target* field with the other 26 fields, in Appendix A, performing as the independent (*In*) variables. See Figure 10 for the Logistic Regression Clementine Stream.

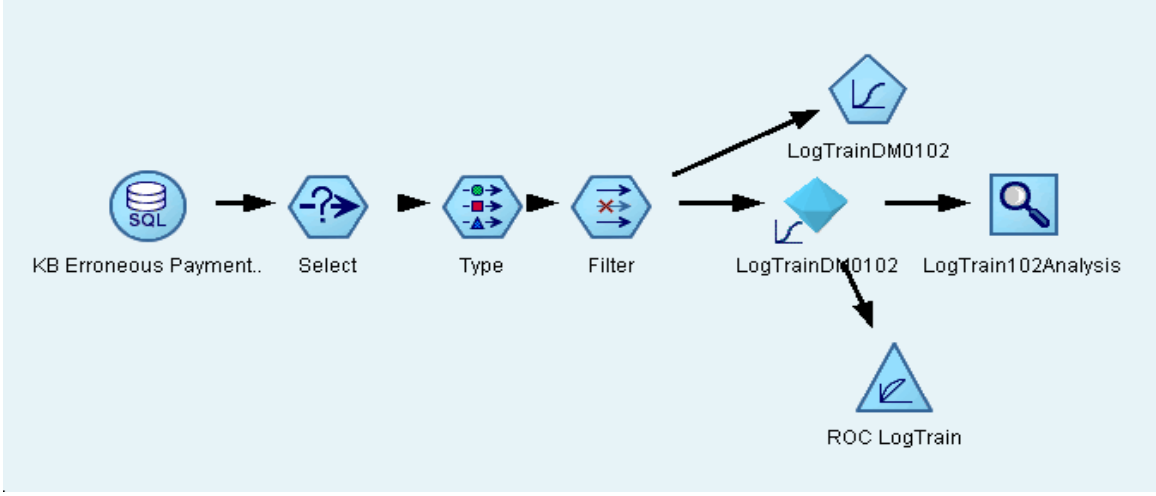



Figure 10. Clementine Logistic Regression Stream

For the Logistic Regression Node, Model options are selected from among Main Effects, Full Factorial and Custom. The Main Effects model includes the input fields individually and does not test interactions between input fields, whereas the Full Factorial includes all two-term interactions. The Full Factorial models are better able to capture complex relationships, but are also much more difficult to interpret and more likely to suffer from

overfitting. [7] The Main Effects model was selected for its ease of model development and understanding.

### 1. Logistic Regression Analysis Node Output

The Analysis Node output in Table 4 shows that the Logistic Regression Node correctly predicts 81.55% of the 3773 records. The percentages for the Coincidence Matrix are displayed as well. The TPF and TNF are fairly high, and will be compared to the C&RT Analysis Node output to determine the tool which is more effective at predicting Erroneous Payments. Only a training set model of the 3773 records will be generated for comparison to the C&RT model.

**Logistic Regression**

| Predicted Probability with Target value | | |
|---|---|---|
| Correct | 3077 | 81.55% |
| Wrong | 696 | 18.45% |
| Total | 3773 | |

| Coincidence Matrix | | |
|---|---|---|
| | 1 | 0 |
| 1 | 414 | 530 |
| 0 | 166 | 2663 |

| Coincidence Matrix | | | |
|---|---|---|---|
| | Actual | | |
| | | EP* (1) | NEP* (0) |
| Predicted | 1 | TPF 71.38% | FPF 16.60% |
| | 0 | FNF 28.62% | TNF 83.40% |

*note: EP is Erroneous Payment/NEP is Non-Erroneous Payment

Table 3. Analysis of DM0102 Training Set for Logistic Regression

## D. CLASSIFICATION AND REGRESSION TREES (C&RT)

The Clementine software package offers both a Classification and Regression Tree Node and a C5.0 algorithm that can build a decision tree or a ruleset. I have chosen to use the C5.0 algorithm, because of its good performance. The C5.0 model tends to be easier to understand than some other model types, since the rules derived from the model have a very straightforward interpretation.

The decision tree is a description of the splits found by the algorithm. Each terminal or "leaf" node describes a particular subset of the training data and each case in the training data belongs to exactly one terminal node in the tree. In other words, exactly one prediction is possible for any particular data record presented to a decision tree. [7]

The ruleset is a set of rules that tries to make predictions for individual records. Rulesets are derived from decision trees and in a way represent a simplified or distilled version of the information found in the decision tree. Rulesets can retain most of the important information from a full decision tree, but with a less complex model. Because of the way rulesets work, they do not have the same properties as decision trees. The most important difference is that with a ruleset, more than one rule may apply for any particular record or no rules at all may apply. If multiple rules apply, each rule gets a weighted "vote" based on the confidence associated with that rule and the final prediction is decided by combining the weighted votes of all of the rules that apply to the

record in question. If no rule applies, a default prediction is assigned to the record. [7]

The ruleset provides the analyst and the decision maker the opportunity to look at a generated ruleset in total and determine which part adds more value to the analysis. This would help give some insight into trends within the data. Figure 12 shows the Clementine C5.0 Stream. Appendix C shows the generated ruleset.

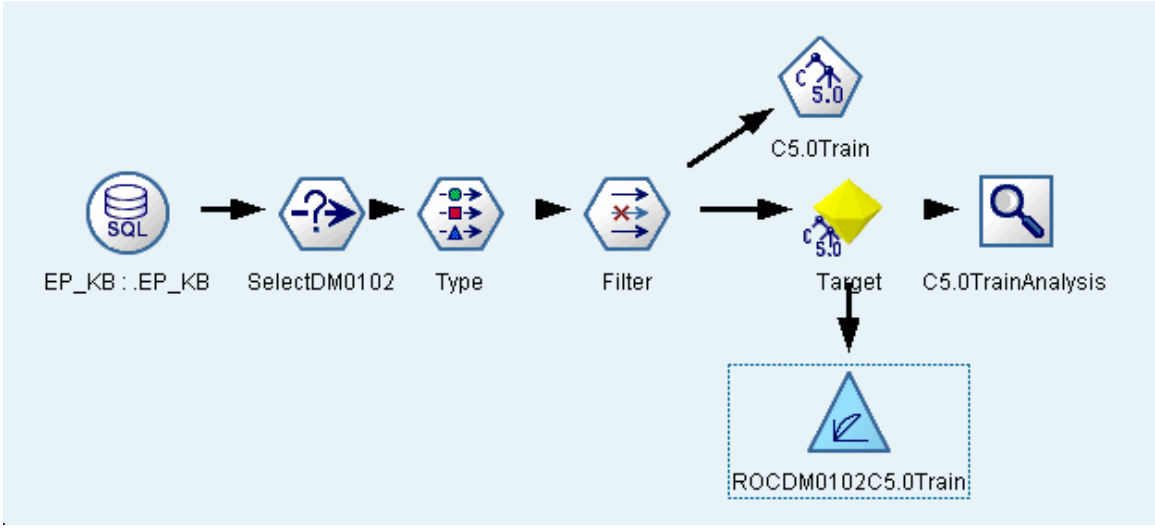



Figure 11. Clementine C5.0 Stream

### 1. C5.0 Training Set Analysis Node Output

This initial Training Set is generated for comparing to the Logistic Regression model. Once a modeling tool is selected a Training and Test Set is generated to check the models ability to predict. The Analysis Node output shows that the C5.0 Node will correctly predict 93.56% of the 3773 records. The pruning severity and node splitting were selected to maximize the predictive ability of the C5.0 stream for this data set. The values selected for the best pruning severity and node split mix were 75% and 3,

respectively. The percentages for the Coincidence Matrix are displayed in Figure 13.

**C5.0**

| Predicted Probability with Target value | | |
|---|---|---|
| Correct | 3530 | 93.56% |
| Wrong | 243 | 6.44% |
| Total | 3773 | |

| Coincidence Matrix | | |
|---|---|---|
| | 1 | 0 |
| 1 | 764 | 180 |
| 0 | 65 | 2764 |

| Coincidence Matrix | | |
|---|---|---|
| Actual | | |
| | EP* (1) | NEP* (0) |
| Predicted 1 | TPF 92.16% | FPF 6.11% |
| Predicted 0 | FNF 7.84% | TNF 93.89% |

*note: EP is Erroneous Payment/NEP is Non-Erroneous Payment

Table 4. C5.0 Analysis Node Output

## E. C5.0 AND LOGISTIC REGRESSION MODEL COMPARISON

Taking the data from Sections B and C, a side by side comparison will be performed to determine the best model to predict the dependent, Target field variable. From Table 5, one can see that the C5.0 is preferred with a correct classification rate of 93.56% over an 81.55% predicted probability for Logistic Regression. In addition, is that the TPF value is clearly higher in the C5.0 Coincidence Matrix over the Logistic Regression. The FPF and FNF are smaller which shows that the numbers of false predictions

would be lower with this model. Clearly the C5.0 is the better model for prediction in this case.

**C5.0**

| Comparing Predicted Probability with Target value | | |
|---|---|---|
| Correct | 3530 | 93.56% |
| Wrong | 243 | 6.44% |
| Total | 3773 | |

| Coincidence Matrix | | |
|---|---|---|
| | 1 | 0 |
| 1 | 764 | 180 |
| 0 | 65 | 2764 |

| Coincidence Matrix | | | |
|---|---|---|---|
| | Actual | | |
| | | EP* (1) | NEP* (0) |
| Predicted | 1 | TPF 92.16% | FPF 6.11% |
| | 0 | FNF 7.84% | TNF 93.89% |

**Logistic Regression**

| Comparing Predicted Probability with Target value | | |
|---|---|---|
| Correct | 3077 | 81.55% |
| Wrong | 696 | 18.45% |
| Total | 3773 | |

| Coincidence Matrix | | |
|---|---|---|
| | 1 | 0 |
| 1 | 414 | 530 |
| 0 | 166 | 2663 |

| Coincidence Matrix | | | |
|---|---|---|---|
| | Actual | | |
| | | EP* (1) | NEP* (0) |
| Predicted | 1 | TPF 71.38% | FPF 16.60% |
| | 0 | FNF 28.62% | TNF 83.40% |

*note: EP is Erroneous Payment/NEP is No Erroneous Payment

Table 5. Coincidence Matrices Comparison

In addition to the previous information, the ROC curves in Figure 12 enhance the Coincidence Matrix analysis, by showing that the curve on the left, (C5.0), has a larger area underneath the upper curve. As discussed in Chapter III, section 4(b), the closer the upper curve is to the straight line, the stronger the evidence that the TPF and FNF are almost equal. This is not a desirable characteristic for an ROC curve. The further these lines are apart the better. The one that performs better is used to determine the cutoff point for the predictive probabilities. The curve on the left (C5.0) is the better one in this case.

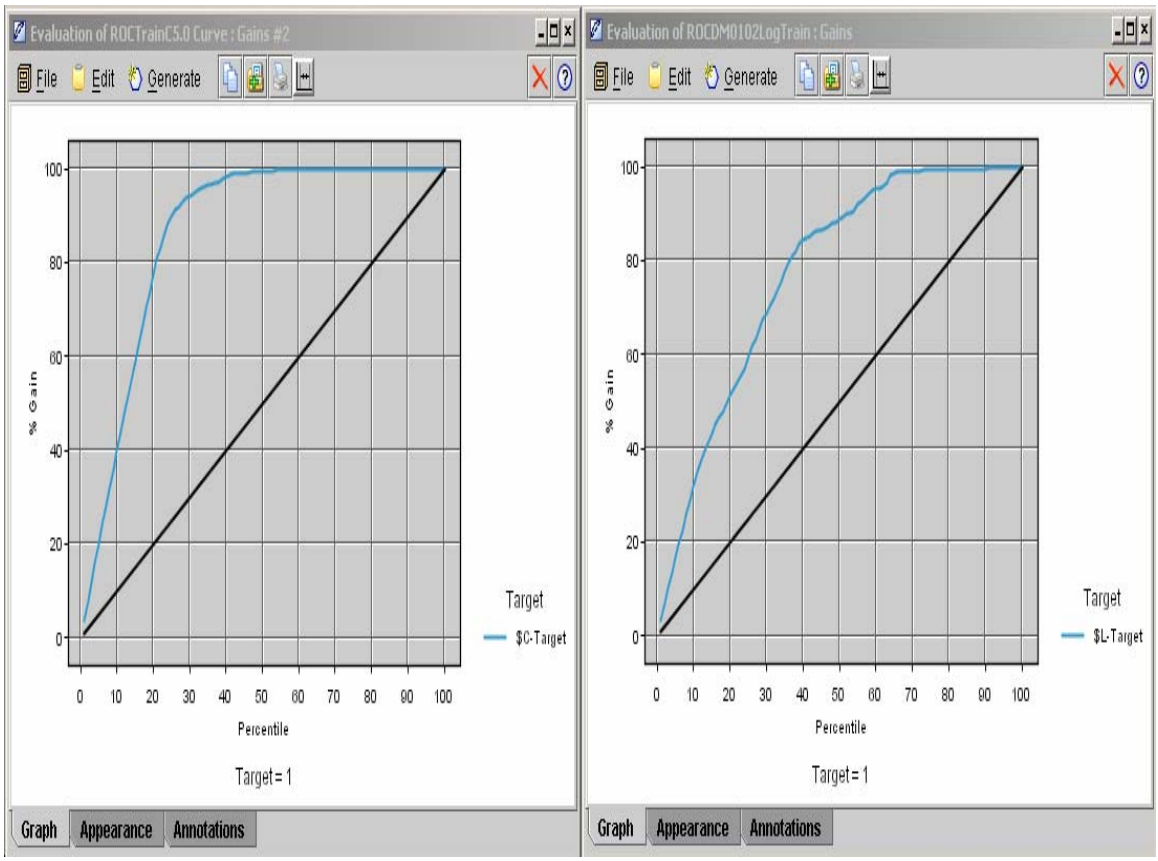



Figure 12. ROC Curve Comparison

## F. C5.0 PERFORMANCE ANALYSIS

The Training and Test Stream is shown in Figure 13. This stream generates a C5.0 Train and C5.0 Test output which will be used to determine if the pruning severity and node split works well on a test set of the overall dataset. The parameter values of best pruning severity equal to 75% and node split equal to 3 were determined for the training set, and will be applied to both the Train and Test data.

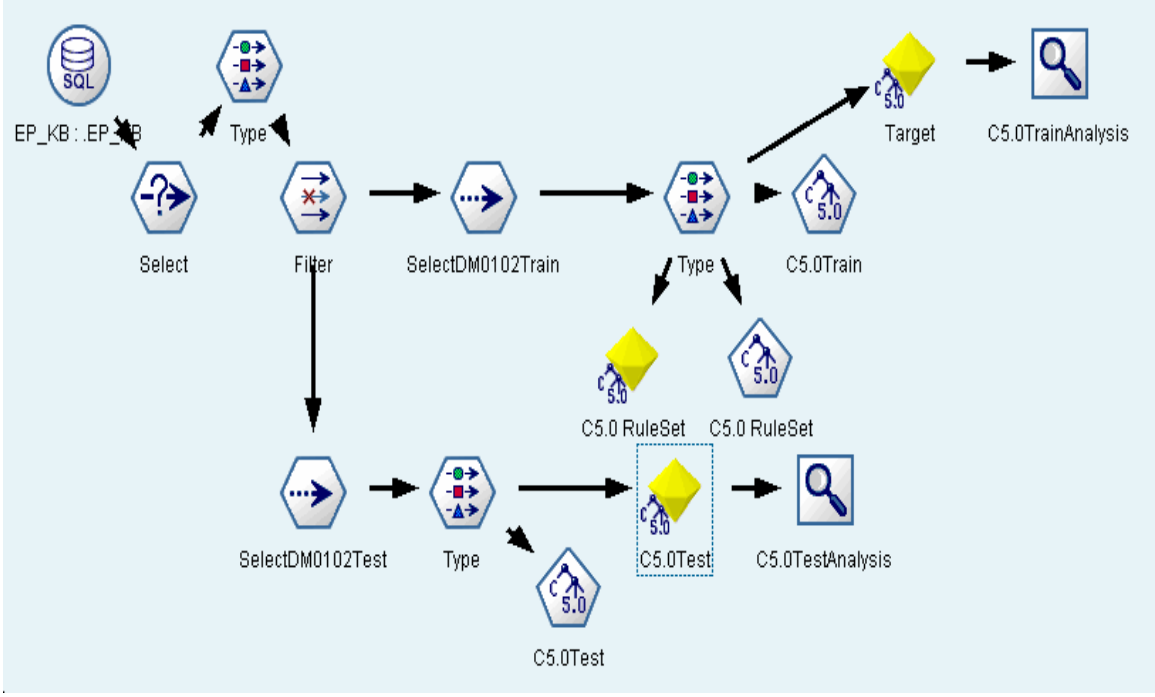
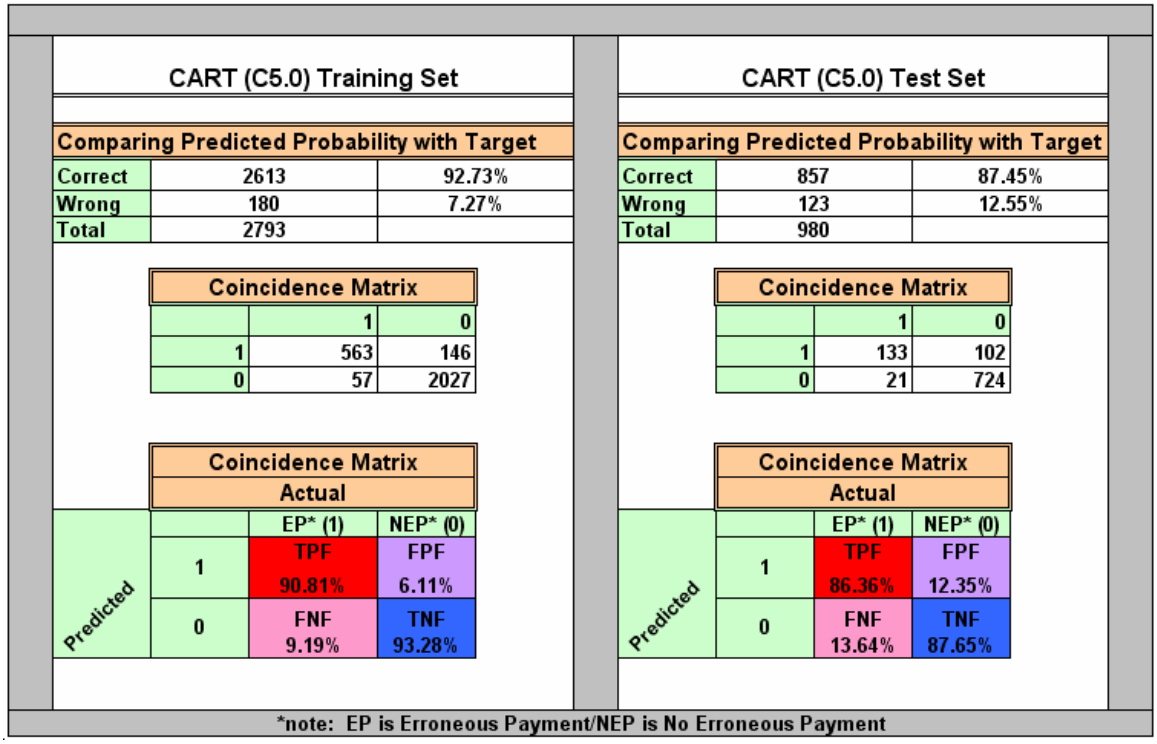



Figure 13. C5.0 Training and Test Stream

The Training and Test Sets are generated randomly with a 75/25 split and the C5.0 node run with a 75/3 pruning severity and node split. The results for the Training and Test Sets are shown in Table 6. The results show predicted probability of 92.73 % and 87.45 % for the Training and Test Set, respectively, a difference of approximately 5 %. The model performed well.



**CART (C5.0) Training Set**

| Comparing Predicted Probability with Target | | |
|---|---|---|
| Correct | 2613 | 92.73% |
| Wrong | 180 | 7.27% |
| Total | 2793 | |

| Coincidence Matrix | | |
|---|---|---|
| | 1 | 0 |
| 1 | 563 | 146 |
| 0 | 57 | 2027 |

| Coincidence Matrix | | | |
|---|---|---|---|
| Actual | | | |
| | | EP* (1) | NEP* (0) |
| Predicted | 1 | TPF 90.81% | FPF 6.11% |
| | 0 | FNF 9.19% | TNF 93.28% |

**CART (C5.0) Test Set**

| Comparing Predicted Probability with Target | | |
|---|---|---|
| Correct | 857 | 87.45% |
| Wrong | 123 | 12.55% |
| Total | 980 | |

| Coincidence Matrix | | |
|---|---|---|
| | 1 | 0 |
| 1 | 133 | 102 |
| 0 | 21 | 724 |

| Coincidence Matrix | | | |
|---|---|---|---|
| Actual | | | |
| | | EP* (1) | NEP* (0) |
| Predicted | 1 | TPF 86.36% | FPF 12.35% |
| | 0 | FNF 13.64% | TNF 87.65% |

*note: EP is Erroneous Payment/NEP is No Erroneous Payment

Table 6. Comparison of C5.0 Training and Test Set Data

# VI. CONCLUSIONS AND RECOMMENDATIONS

## A. CONCLUSION

The goal of this thesis was to look at the Knowledge Base of Erroneous Payments (KBOEP) and generate models for duplicate payments. These models were then compared and the best one was selected based on its predictive ability. The C5.0 algorithm provided the best results and the most flexibility in generating a ruleset. Because of the continual working environment of the auditing process the model's ruleset may change as the KBOEP grows and more analysis will have to be done to see if the model's predictive ability changes as well.

DFAS has made positive contributions to the auditing of payments for goods and services within the DOD. It is clear that the administrative portion of auditing vendor payments is a critical piece of performing our role as good stewards of the taxpayers' dollars.

## B. RECOMMENDATIONS

In order to validate the usefulness of the best model selected, DFAS IR should employ this model in parallel with their current manual record selection process. The process should compare manually selected records to electronic records and an analysis done on which technique performed better. The on-site audits will validate the successfulness of the models capturing the right records for erroneous payments.

We recommend using more C&RT techniques on the master database from which the KBEOP was drawn to see if new models can be developed to validate or enhance the current ones used for detecting erroneous payments. The master

database should be analyzed using both supervised and unsupervised algorithms to determine the validity of current fraud detection models and to gain insight into new fraudulent behaviors.

A final area to look at would be to run the Logistic Regression and C&RT algorithms to determine and develop models for the DM0109, DM0110, DM0111 and DM0210 fraud detection models, in addition to the master database.

# APPENDIX A – FIELD NAMES

| Field Name | Field Abbreviation | Type | Values |
|---|---|---|---|
| Target | Target | Flag | 1/0 |
| Number In Group | NumInGrp | Numeric | [2,3,4,5,7] |
| Invoice Date Difference | InvDtDiff | Numeric | [0,610] |
| Invoice Received Difference | InvRcvdDiff | Numeric | [0,439] |
| Merchandise Accepted Date Difference | MdseAccDtDiff | Numeric | [0,610] |
| Merchandise Delivered Date Difference | MdseDelDtDiff | Numeric | [0,621] |
| Check Date Difference | ChkDtDiff | Numeric | [0,609] |
| Payment Method Difference | PmtMethDiff | Flag | 1/0 |
| Manual Indicated Difference | ManIndDiff | Flag | 1/0 |
| Electronic Fund Transfer Accounting Difference | EFT_AcctDiff | Flag | 1/0 |
| Electronic Fund Transfer Return Difference | EFT_RtnDiff | Flag | 1/0 |
| Tax Identification Number Difference | TINDiff | Flag | 1/0 |
| Remit to Difference | Rmt_ToDiff | Flag | 1/0 |
| Remit Line 1 Difference | Rmt_L1Diff | Flag | 1/0 |
| Remit Line 2 Difference | Rmt_L2Diff | Flag | 1/0 |
| Remit City Difference | Rmt_CityDiff | Flag | 1/0 |
| Remit Zip Difference | Rmt_ZipDiff | Flag | 1/0 |
| Maximum Invoice Received versus Invoice date | MaxInvRcvdvsInv_dt | Numeric | [1,2245] |
| Minimum Invoice Received versus Invoice date | MinInvRcvdvsInv_dt | Numeric | [1,2245] |
| Appropriation Identification Difference | Appr_IDDiff | Flag | 1/0 |
| Appropriation Fiscal Year Difference | Appr_FYDiff | Flag | 1/0 |
| Appropriation Limit Difference | Appr_LimtDiff | Flag | 1/0 |
| Line of Accounting Difference | LoaDiff | Flag | 1/0 |
| Reissue Rejected | Reissue_Reject | Flag | 1/0 |
| Electronic Fund Transfer Rejected | EFTRej | Flag | 1/0 |
| Manual Payment | Man_Pymt | Flag | 1/0 |
| Maximum Invoice Amount | Max_INV_AMT | Numeric | [200.00, 1,847,128.00] |

Table 7. Field Names

THIS PAGE INTENTIONALLY LEFT BLANK

# APPENDIX B - CLEMENTINE DATA STREAM AND NODE DIALOG BOXES

This appendix will explain the Clementine applications and Dialog boxes used in this thesis.

## A. CLEMENTINE LOGISTIC REGRESSION STREAM EXPLANATION

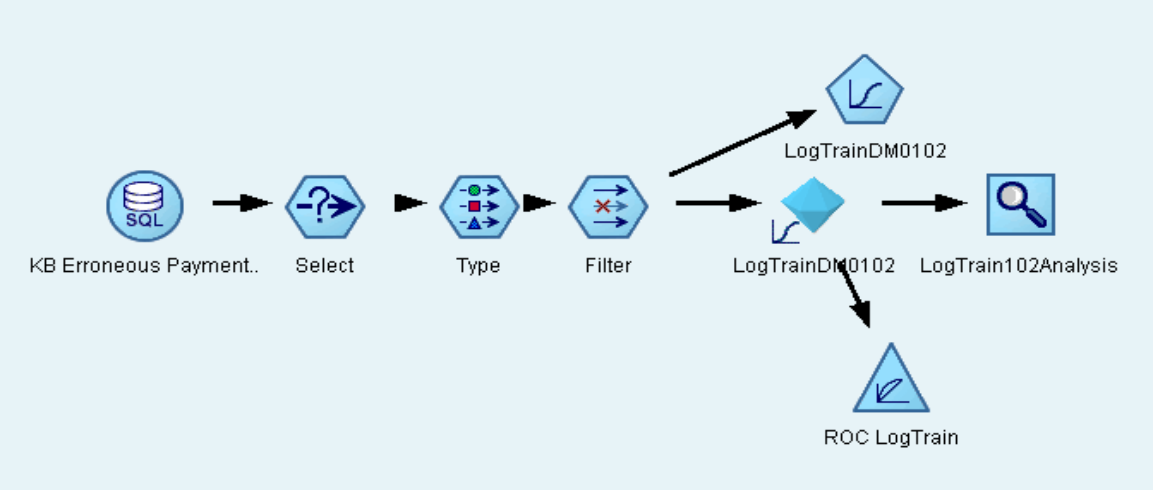



Figure 14. Clementine Logistic Regression Stream

Each icon in Figure 14 is known as a node. Dialog boxes are generated for each node in order to change certain features of that node. As each node is explained for a Clementine stream it will not be explained throughout the rest of this Appendix. The Database (SQL), Select, Type and Filter Nodes are the same throughout.

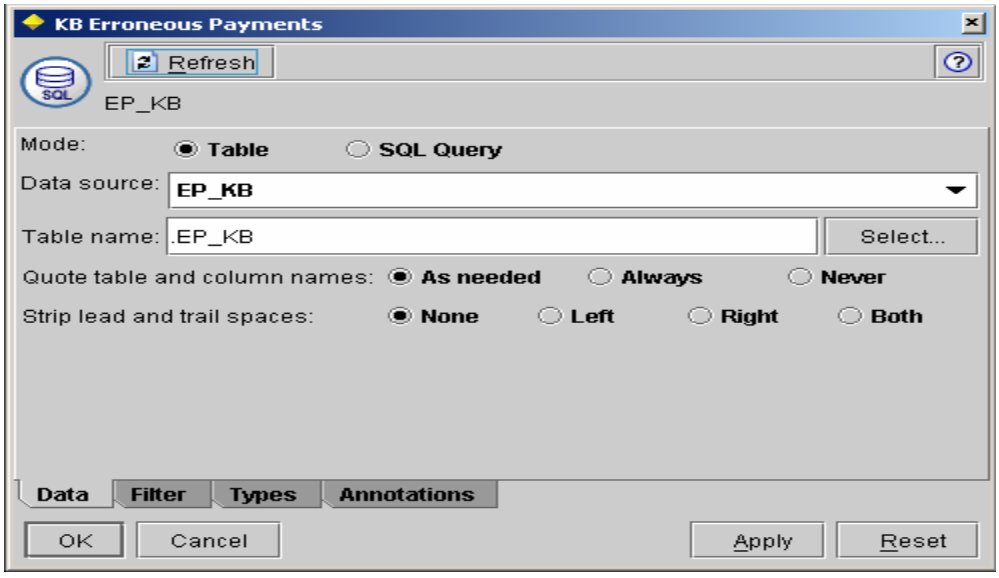



Figure 15. SQL Dialog Box

The Database Node allows the user to import data from a variety of other packages, including Excel, MS Access, Dbase, SAS (NT version only), Oracle and Sybase, using the ODBC source node. The database that was used was in MS Access format.

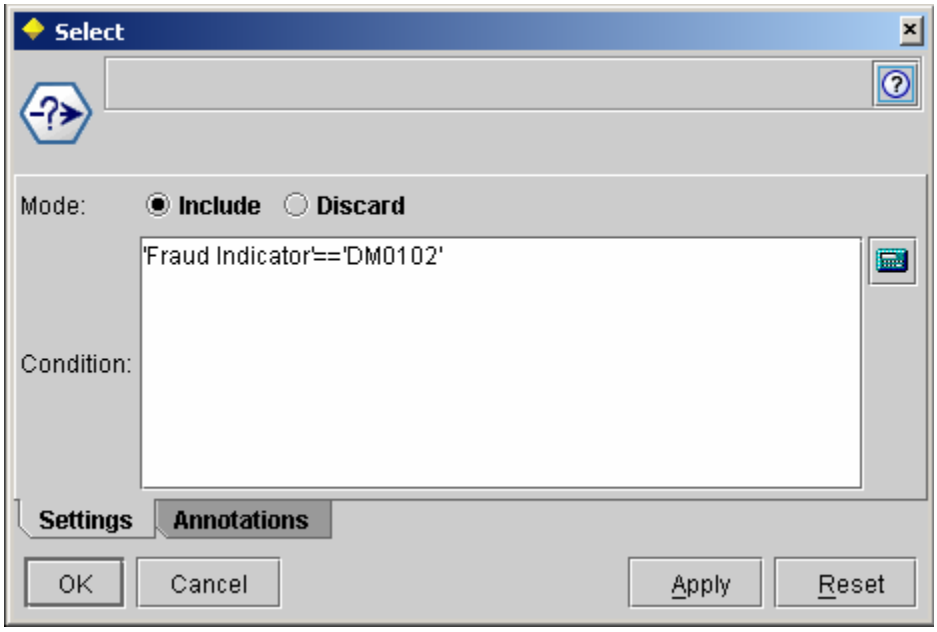



Figure 16. Select Dialog Box

The Select Node allows the user to select or discard a subset of records from the data stream based on a specific condition, such as selecting the DM0102 fraud model from the other five fraud indicators.

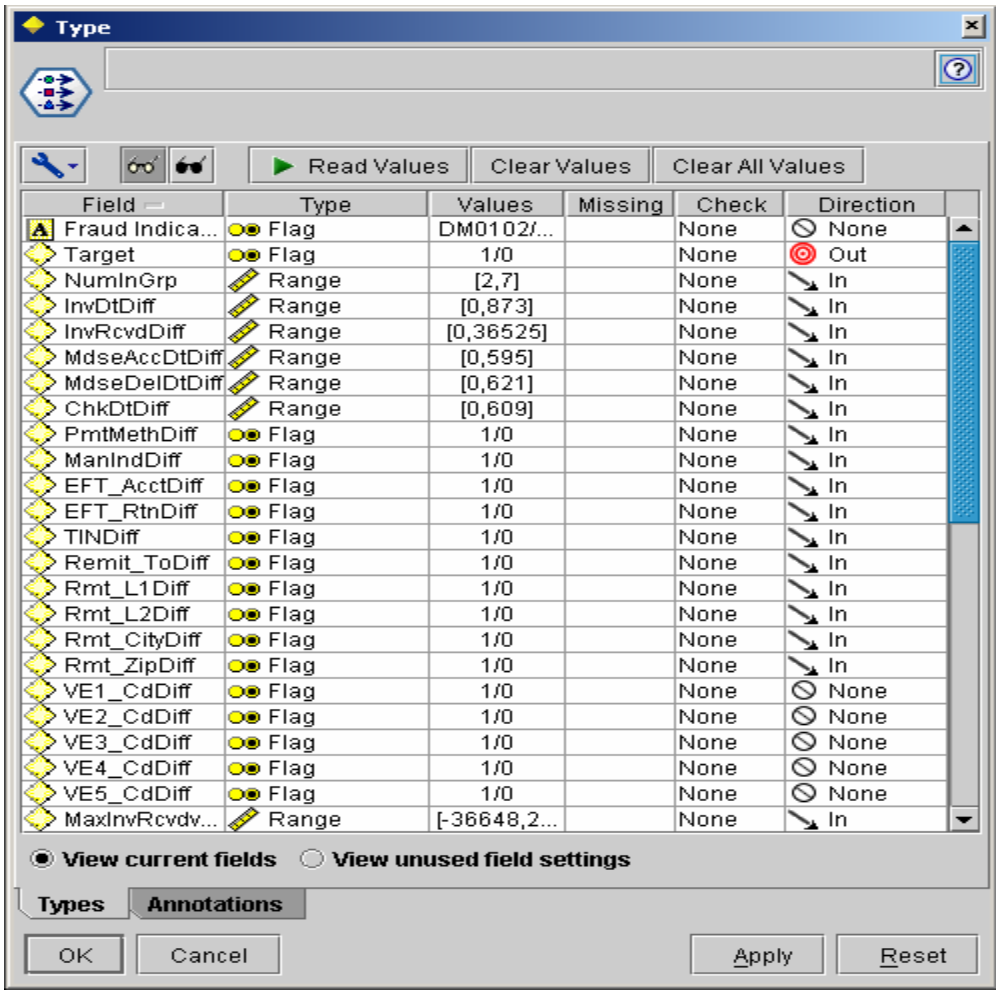



Figure 17. Type Dialog Box

The Type Node can specify a number of field properties. Most important are the Type and Direction fields. The Type field is used to describe characteristics of the data in a given field such as Flag, Range, Discrete, Set or Typeless. The Direction section is used to tell the Modeling Nodes whether fields will be **Input** (predictor fields) or **Output** (predicted fields) for a machine learning process. **Both** and **None** are also available directions.

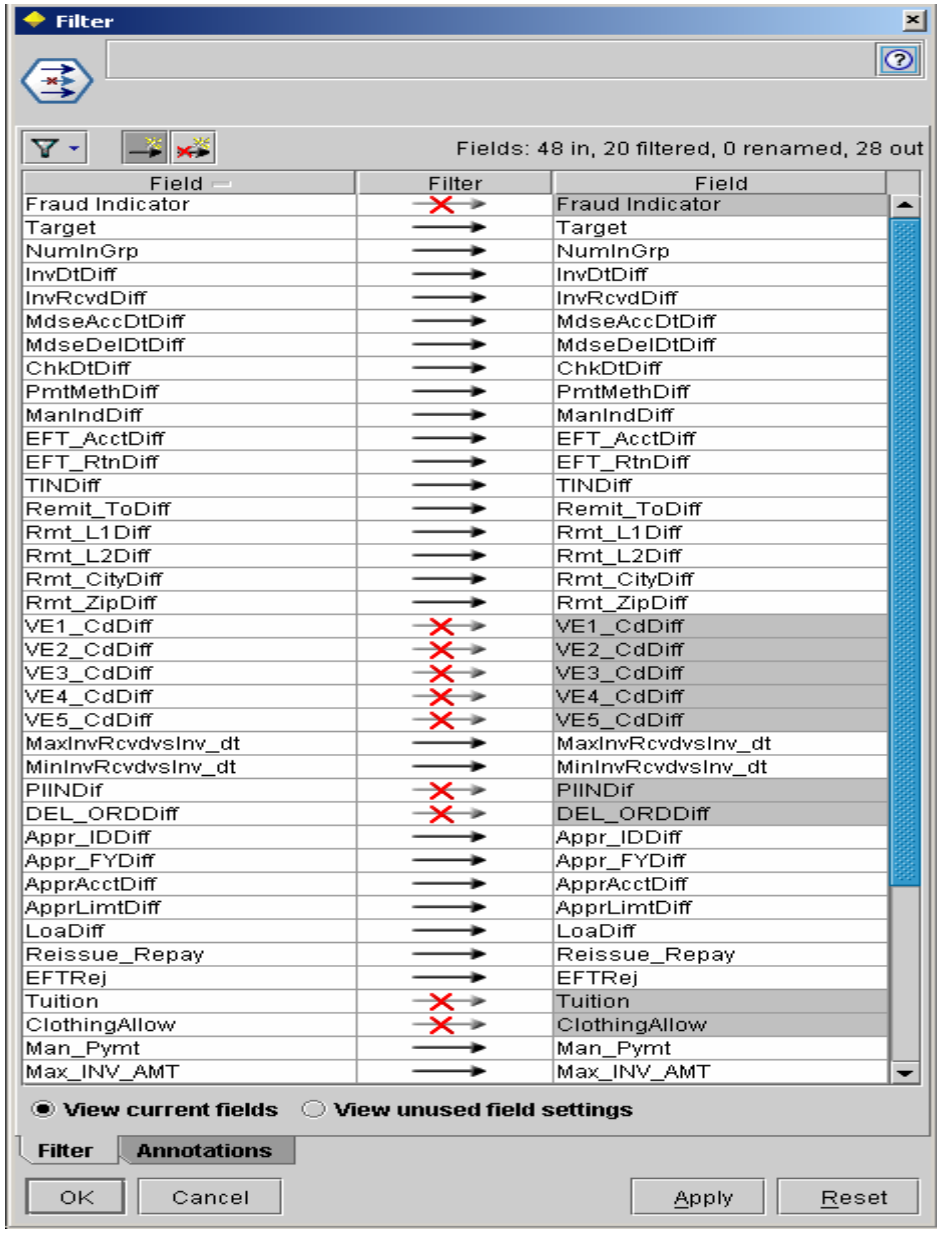



Figure 18. Filter Dialog Box

The Filter Node has three functions: filter (or discard) fields, rename fields and map fields. In this thesis, 20 of 48 fields were filtered, because they were not tied directly to the fields normally seen in a payment document.

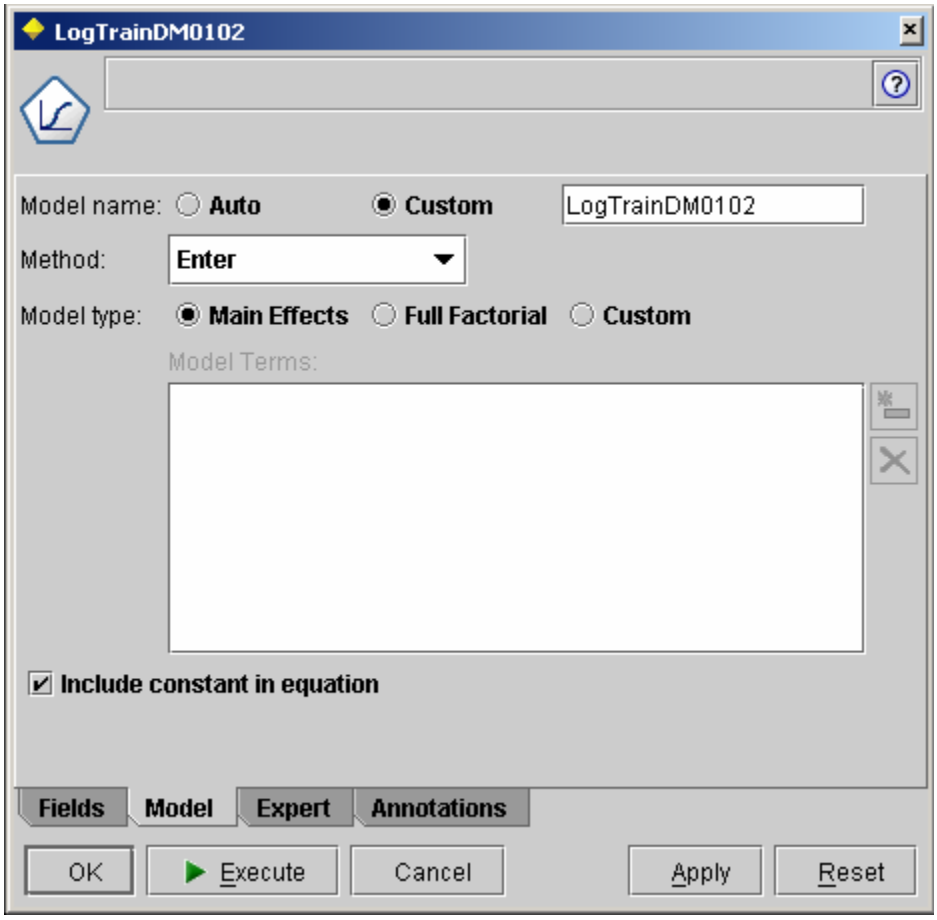



Figure 19. Logistic Regression Dialog Box

The Logistic Regression Node works by building an equation that relates the input field values to the probabilities associated with each of the output field categories. Once the model is generated, it can be used to estimate probabilities for new data. For each record, a probability of membership is computed for each possible output category. The target category with the highest probability is assigned as the predicted output value for that record.

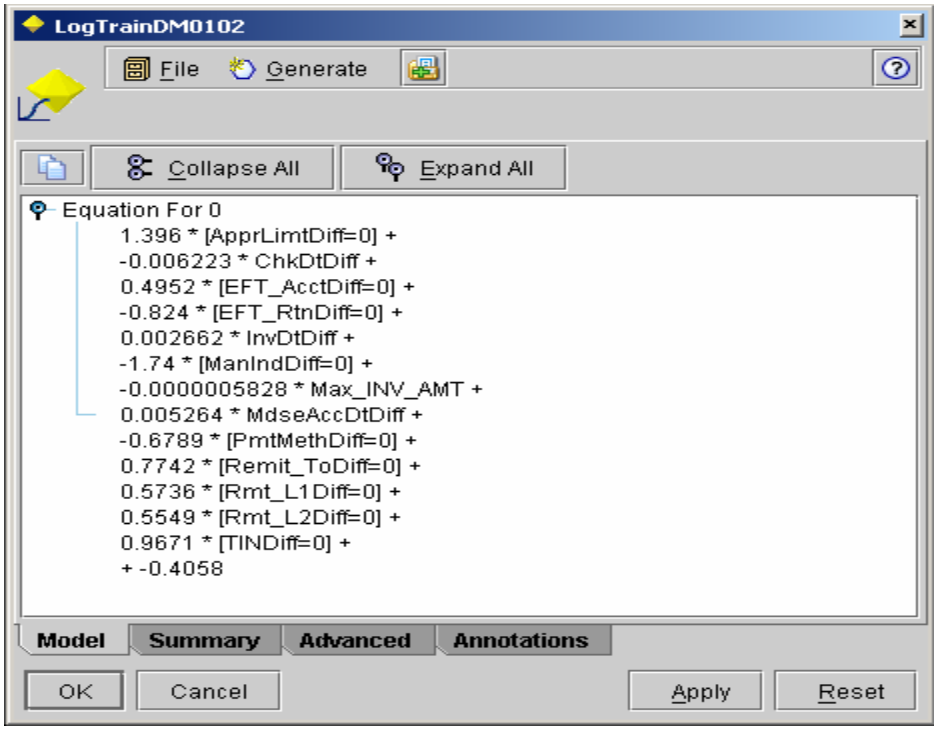



Figure 20. Logistic Regression Model Summary Dialog Box

Figure 20 is an example of the output of the Logistic Regression Node.

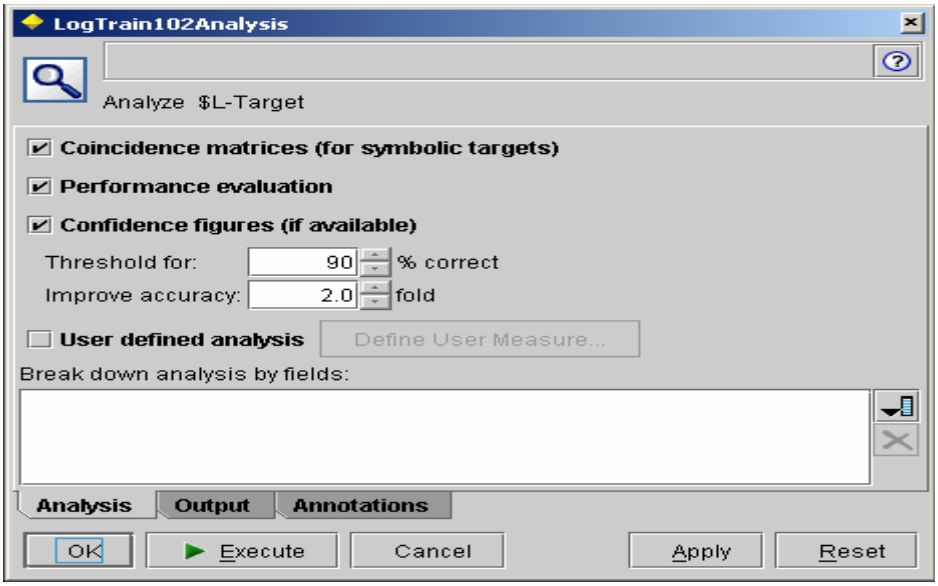



Figure 21. Analysis Dialog Box

The Analysis Dialog Box in Figure 21 allows the user to specify the details of the analysis by generating coincidence matrices, performance evaluations and confidence figures. Figure 22 shows an example of the output that is generated.

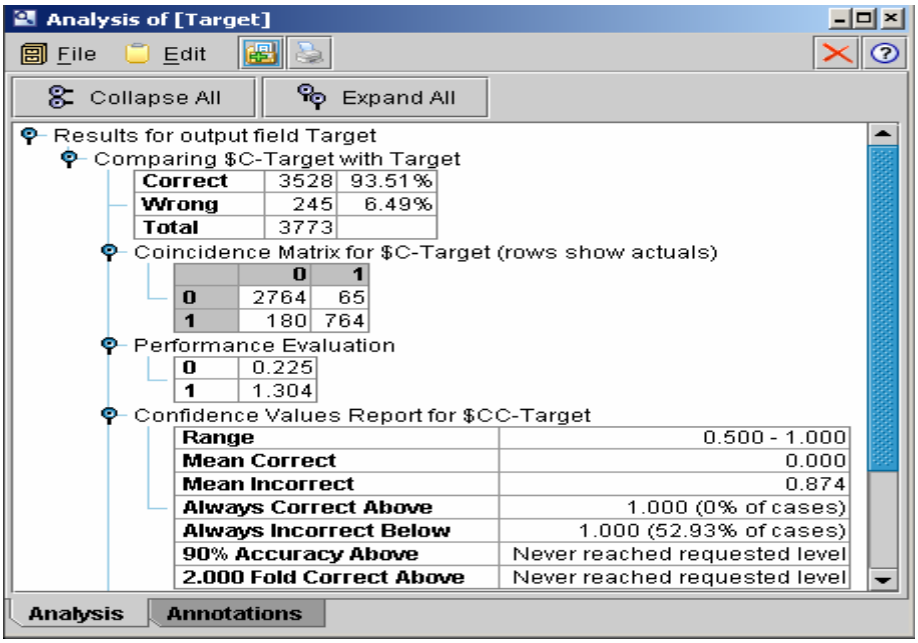



Figure 22. Analysis Output Dialog Box

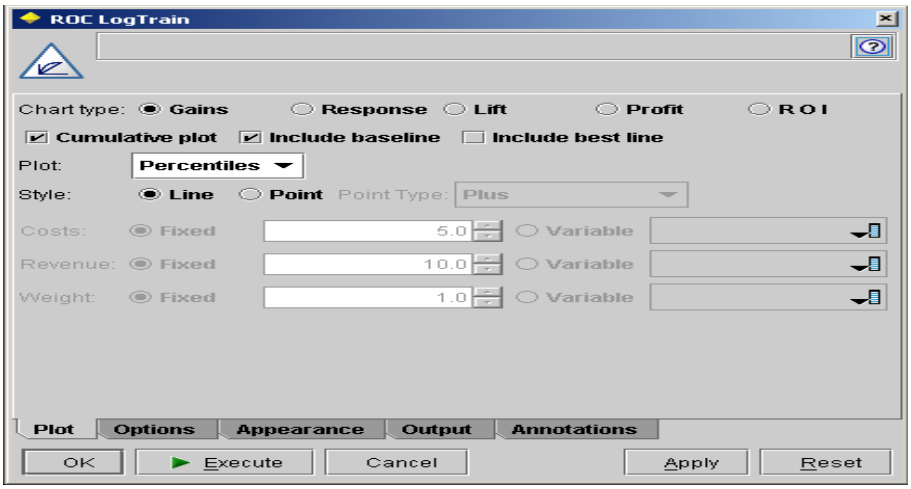



Figure 23. Evaluation Dialog Box

Figure 23 shows the Evaluation Chart Dialog Box that offers an easy way to generate, evaluate and compare predictive models to choose the best model for your application. Evaluation charts show how models perform in predicting particular outcomes. They work by sorting records based on the predicted value and confidence of the prediction, splitting the records into groups of equal size (quantiles) and then plotting the value of the business criterion for each quantile, from highest to lowest. [7] An example of an ROC Curve is shown in Figure 24.

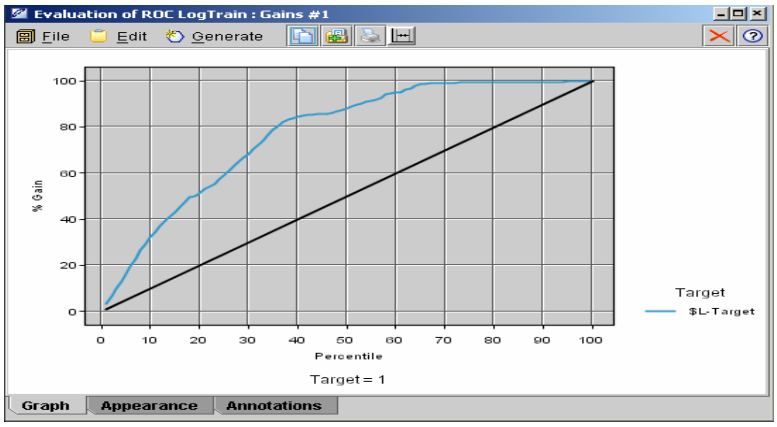



Figure 24. Evaluation Output Dialog Box

## B. C5.0 STREAM EXPLANATION

The C5.0 Node is explained in detail in Chapter IV(C).

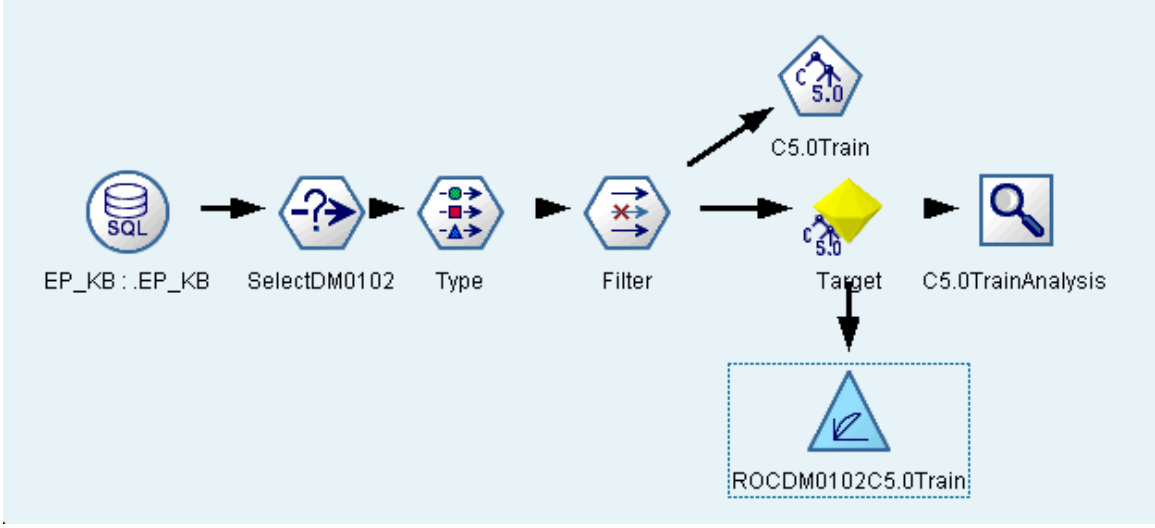



Figure 25. C5.0 Stream

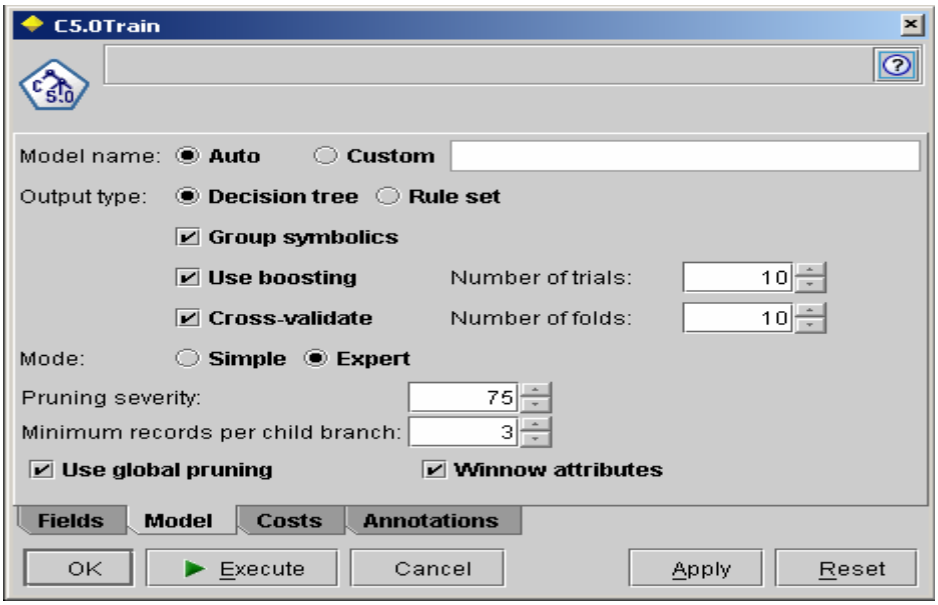



Figure 26. C5.0 Dialog Box

Figure 26 shows the dialog box and the fields where the pruning severity and minimum number of fields to split per branch are selected.

## C. C5.0 TRAIN AND TEST SET STREAM EXPLANATION

Figure 26 shows the stream that examines the strength of the C5.0 model and its ability to predict the outcomes.

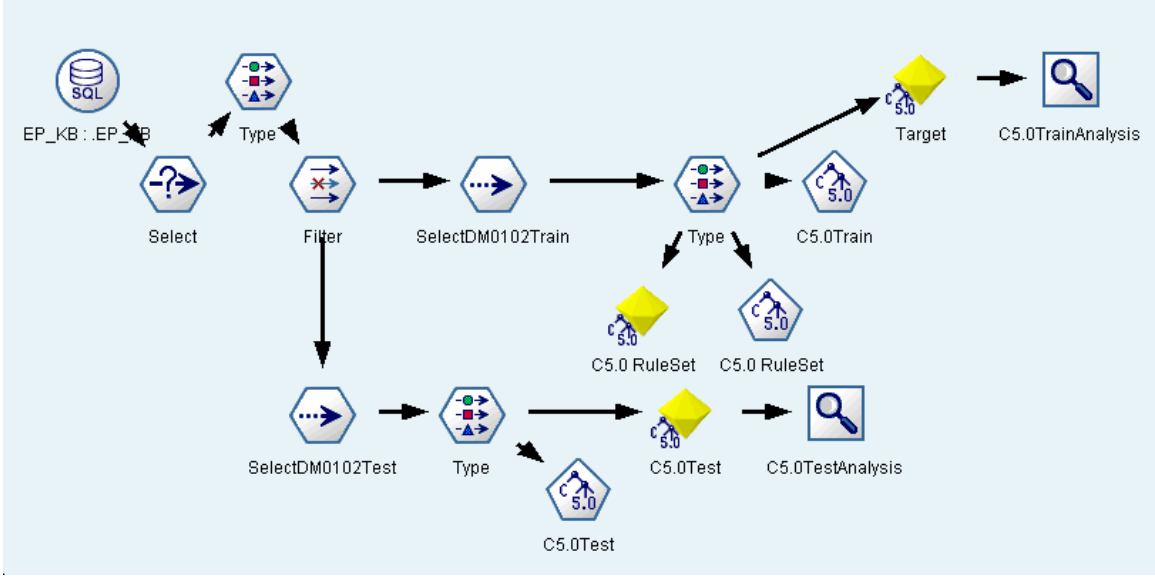



Figure 27. C5.0 Train and Test Stream

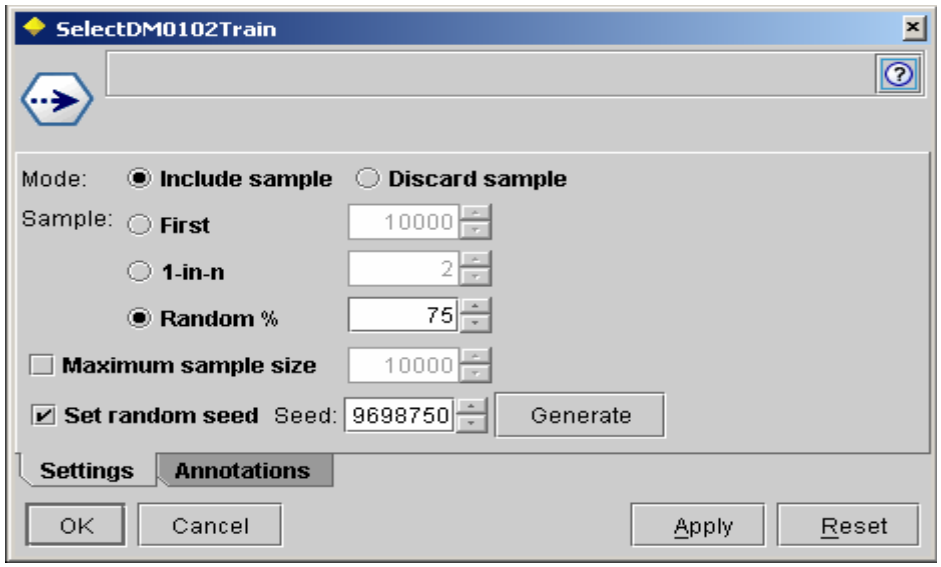



Figure 28. DM0102 Select Dialog Box

Figure 28 shows the Select Node Dialog Box

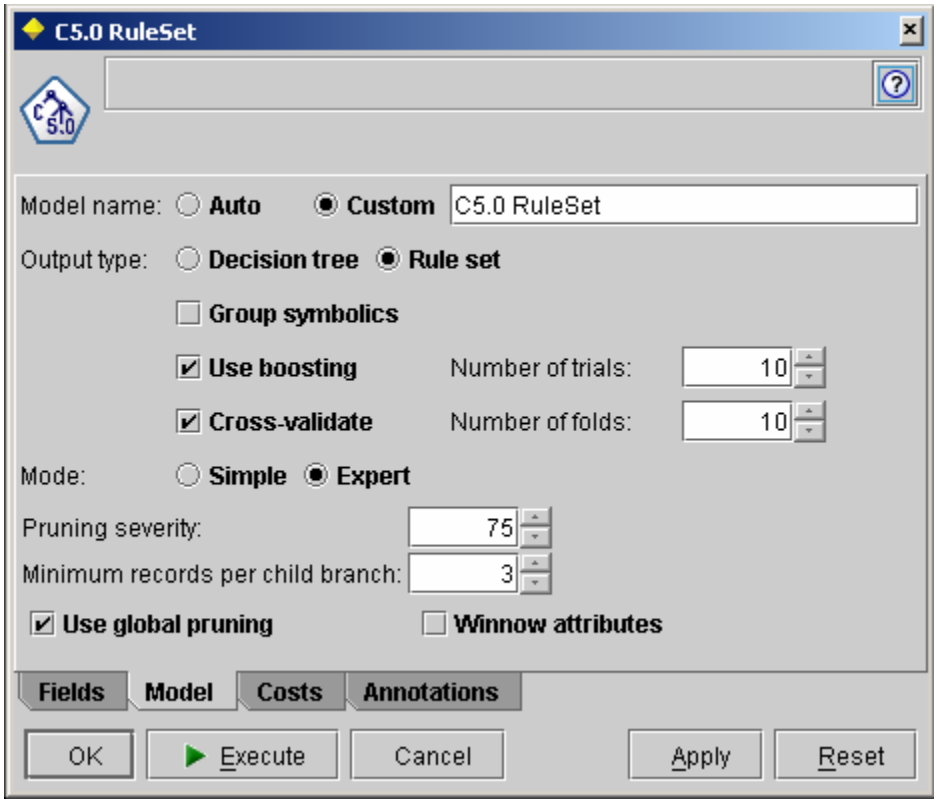



Figure 29. C5.0 Ruleset Dialog Box

Figure 29 is the Dialog Box for the selecting the ruleset parameters within the C5.0 Node. Boosting and Cross validation are explained in detail in the Clementine User's Guide [7].

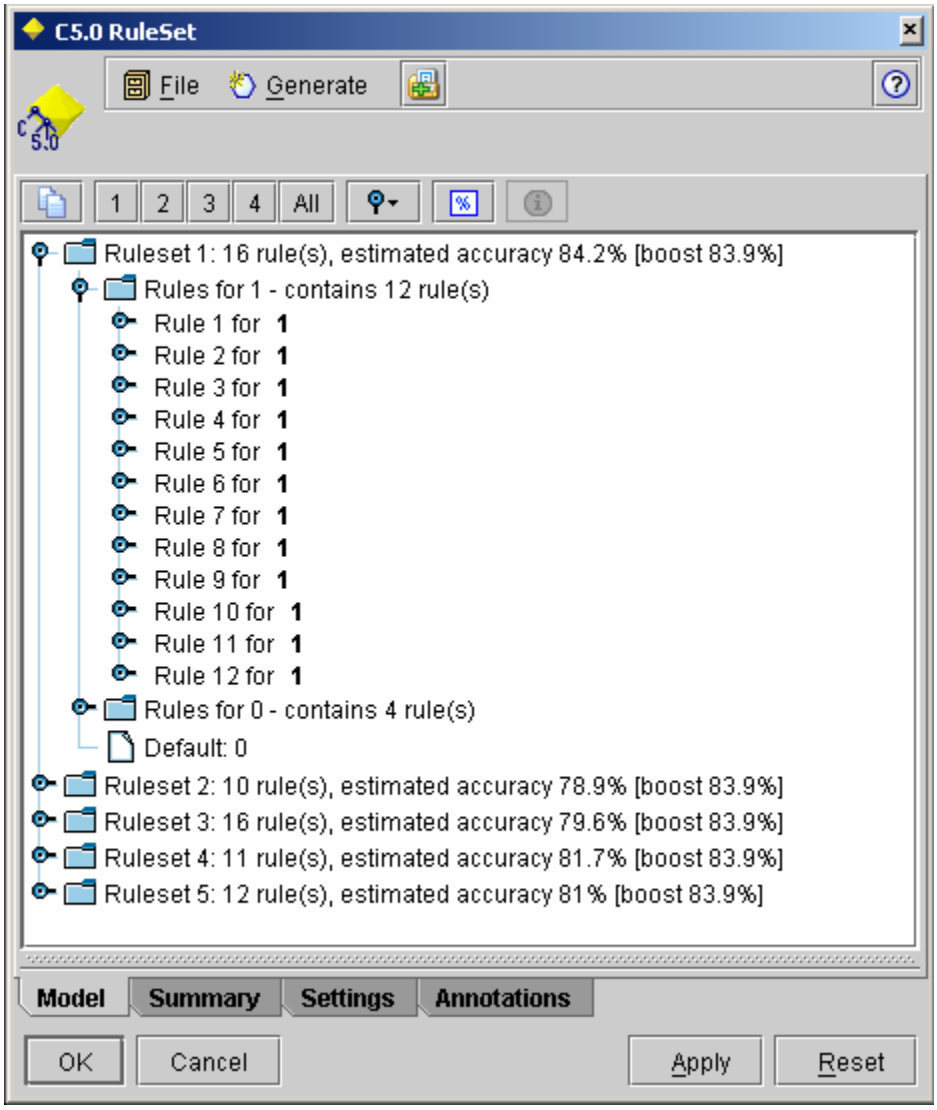



Figure 30. C5.0 Ruleset Output

This is an example of the output generated by selecting Ruleset in the C5.0 Node.

THIS PAGE INTENTIONALLY LEFT BLANK

# APPENDIX C - C5.0 GENERATED RULESET

```
Rules for 1 - contains 13 rule(s)
Rule 1 for 1 (150, 0.901)
      if InvRcvdDiff <= 98
      and TINDiff = 1
      and Max INV AMT > 511.54001
      then 1
Rule 2 for 1 (41, 0.86)
      if MdseAccDtDiff <= 9
      and ChkDtDiff > 0
      and PmtMethDiff = 0
      and EFT AcctDiff = 0
      and EFT RtnDiff = 0
      and MinInvRcvdvsInv dt <= 3
      and Max INV AMT > 741
      then 1
Rule 3 for 1 (11, 0.846)
      if MdseAccDtDiff > 9
      and ChkDtDiff > 0
      and ChkDtDiff <= 3
      then 1
Rule 4 for 1 (4, 0.833)
      if ChkDtDiff > 0
      and EFT AcctDiff = 1
      and EFT RtnDiff = 0
      and TINDiff = 0
      and Max INV AMT <= 1915
      then 1
Rule 5 for  1 (140, 0.824)
      if MdseAccDtDiff <= 9
      and ManIndDiff = 0
      and Rmt L2Diff = 1
      then 1
Rule 6 for  1 (149, 0.815)
      if MdseAccDtDiff <= 9
      and ChkDtDiff > 0
      and PmtMethDiff = 0
      and EFT RtnDiff = 1
      and Rmt L1Diff = 1
      and Appr IDDiff = 0
      then 1
Rule 7 for  1 (40, 0.81)
      if ChkDtDiff > 0
      and ManIndDiff = 0
      and Remit ToDiff = 0
```

```
      and Man Pymt = 1
      then 1
Rule 8 for  1 (34, 0.806)
      if MdseAccDtDiff <= 9
      and ChkDtDiff > 0
      and ManIndDiff = 0
      and EFT RtnDiff = 0
      and TINDiff = 0
      and Remit ToDiff = 1
      then 1
Rule 9 for  1 (3, 0.8)
      if ChkDtDiff > 10
      and ManIndDiff = 1
      and TINDiff = 0
      and Rmt L2Diff = 1
      then 1
Rule 10 for  1 (3, 0.8)
      if ChkDtDiff > 116
      and ManIndDiff = 1
      then 1
Rule 11 for  1 (19, 0.762)
      if InvRcvdDiff <= 0
      and MdseAccDtDiff <= 9
      and ChkDtDiff > 0
      and ManIndDiff = 0
      and EFT RtnDiff = 0
      and Appr FYDiff = 1
      then 1
Rule 12 for  1 (45, 0.745)
      if ChkDtDiff > 0
      and ManIndDiff = 0
      and Man Pymt = 1
      then 1
Rule 13 for  1 (129, 0.695)
      if InvRcvdDiff > 0
      and MdseAccDtDiff <= 9
      and ChkDtDiff > 0
      and EFT AcctDiff = 0
      and TINDiff = 0
      then 1
Rules for 0 - contains 7 rule(s)
Rule 1 for  0 (819, 0.985)
      if ChkDtDiff <= 116
      and ManIndDiff = 1
      and TINDiff = 0
      and Rmt L2Diff = 0
      then 0
```

```
Rule 2 for  0 (846, 0.982)
      if MdseAccDtDiff <= 9
      and PmtMethDiff = 1
      and TINDiff = 0
      and Rmt L2Diff = 0
      then 0
Rule 3 for  0 (42, 0.955)
      if InvRcvdDiff <= 0
      and Remit ToDiff = 0
      and MinInvRcvdvsInv dt <= 3
      and Max INV AMT <= 741
      then 0
Rule 4 for  0 (286, 0.951)
      if EFT RtnDiff = 1
      and TINDiff = 0
      and Rmt L1Diff = 0
      then 0
Rule 5 for  0 (246, 0.923)
      if ChkDtDiff <= 0
      then 0
Rule 6 for  0 (1,042, 0.922)
      if InvRcvdDiff <= 0
      and EFT AcctDiff = 0
      and Remit ToDiff = 0
      and MinInvRcvdvsInv dt > 3
      and Appr FYDiff = 0
      and ApprLimtDiff = 0
      then 0
Rule 7 for  0 (194, 0.867)
      if MdseAccDtDiff > 9
      and ChkDtDiff > 3
      and TINDiff = 0
      and Man Pymt = 0
      then 0
```

THIS PAGE INTENTIONALLY LEFT BLANK

# LIST OF REFERENCES


1. Jones-Oxendine, Shawn R., *An Analysis Of DOD Fraudulent Vendor Payments,* Master's Thesis, Naval Postgraduate School, Monterey, California, September 1999.

2. Jenkins, Donald J., Evaluation of Fraud Detection Data Mining Used in the Auditing Process of the Defense Finance and Accounting Service, Master's Thesis, Naval Postgraduate School, Monterey, California, June 2002.

3. DFAS IR Seaside to DFAS Headquarters Memorandum, Second Quarter Progress Report, March 2002.

4. U. S. General Accounting Office, GAO Report GAO-02-069G, *Strategies to Manage Improper Payments: Learning from Public and Private Sector Organizations*, US General Accounting Office, Washington, D.C., October 2001.

5. U. S. General Accounting Office, GAO Report GAO-01-703G, *Strategies To Manage Improper Payments Learning From Public And Private Sector Organizations*, US General Accounting Office, Washington, D.C., May 2001.

6. Defense Finance and Accounting Service, *Improper Payments Data Mining Support:*Final *Report 831-583-3002,* Federal Data Corporation, Contract N00244-96-D-8055, 1999.

7. *Clementine 7.0 Users Guide*, SPSS Inc. Chicago, Illinois. 2002.

8. Rouillard, Gregory W., An Improved Unsupervised Modeling Methodology for Detecting Fraud in Vendor Payment Transactions, Master's Thesis, Naval Postgraduate School, Monterey, California, June 2003.

9. Davia, Howard, *Fraud 101*, John Wiley and Sons, Inc., New York, New York, 2000.

10. DFAS Internal Review, Duplicate Vendor Payments at DFAS-Columbus, April – December 2001.

11. DFAS Internal Review, Vendor Pay Erroneous Payment Audit DFAS-Pacific, December 2002 to March 2003.

12. DFAS Internal Review, Vendor Pay Erroneous Payment Audit DFAS-Charleston, June - December 2003.

13. DFAS Internal Review, Vendor Pay Erroneous Payments Audit DFAS-Kansas City, March - June 2003

14. DFAS Internal Review, FY 02 Erroneous Contract Payments DFAS-Columbus, October 2001 – September.

15. Hamilton, Lawrence C, *Regression with Graphics; A Second Course in Applied Statistics*, Wadsworth, Inc., Belmont, California, 1992.

16. Devore, Jay L., *Probability and Statistics for Engineering and the Sciences*, Duxbury, Thomson Learning, Inc., Fifth Edition, 2000.

17. TIME, *Anderson: The Whistle Not Blown*, http://www.time.com/time/nation/article/0,8599,194573,00.html, July 2004.

18. Hand, David, Heikki Mannila and Padhraic Smyth, *Principles of Data Mining*, The MIT Press, Cambridge, Massachusetts and London England, 2001.

19. The Magnificent ROC, (Receiver Operator Characteristic Curves), www.anesthetist.com/mnm/stats/roc/, August 2004

20. Hosmer, David W., Stanley Lemeshow, *Applied Logistic Regression*, Wiley-Interscience publication, John Wiley & Sons, Inc., 1989.

# INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Internal Review Seaside (Operation Mongoose)
   DOD Center Monterey Bay
   Seaside, California