Tyler Anderson  
DSC-630  
Case Study

**Improper Payment Detection in Department of Defense Financial Transactions**

**Data Understanding: Defining the Data**

This case study describes ongoing work to aid examiners in the detection of illegal, improper, or unusual transactions conducted against the Department of Defense (DoD) financial assets. The main goals of this project are to improve the Defense Finance and Accounting Service's (DFAS) ability to detect these payments and to reduce the manpower required to research them. By taking advantage of current data mining tools, DFAS hopes to enhance the precision of their predictions.

The Defense Manpower Data Center (DMDC) receives an extract from each system then stores them in a Common Data Format (CDF) to standardize many of the data elements. DFAS extracted actual transactions for some of the cases and used source documents to recreate the remaining transactions to appear as they would have in the CDF format. The result was a data set of fraudulent payment candidates that DFAS used to develop models predicting similar transactions. Because many of the transactions were older than the data history, some data reconstructions were incomplete. For example, some transactions were missing data on payment type or payment method. Fortunately for the taxpayers, the number of known fraudulent transactions are still small. However, this was the challenge for the Data Mining effort: trying to predict suspicious payments using a very small set of known fraudulent payments relative to a larger population of non-fraudulent payments.

**Data Preparation**

In this case study the effort was made to identify transformations common to both the known fraudulent payment data and the CDF data. Experts in identifying vendor payment fraud hypothesized dozens of potentially useful transformations of known information that might be useful indicators of fraud. Also transformations that required information not available in the known fraudulent payment data, but available in the CDF data, were included in the data set to be used in future unsupervised learning modeling stages of the project.  
Examples of data transformations made in the first phase included setting flags that identified:

* Payments addressed to P.O. Box or Suite;
* Invoices from the same vendor paid to multiple addresses;
* Invoices from multiple vendors paid to the same address;
* Invoices from the same vendor were not sequential based on date submitted.

Rather than using a single fraud/not-fraud binary label for their output variable, four fraudulent payment types, called types A, B, C, and D, were identified as comprising the different styles of payments in the known fraud data. Although separating the known fraudulent payments into types was not essential to the modeling process, researchers of this case study did believe that the additional information would aid classifiers in identifying suspicious payments and reduce false alarms. The primary reason for this belief was that the types of known fraudulent payments had significantly different behaviors, which would cause a classification algorithm to try to segment the fraudulent payment data into groups anyway. By creating the payment types beforehand, simpler models were more likely to be formed.

**Modeling**

A template Clementine stream containing nodes for reading data, assigning data types, performing simple data transformations, and displaying model results was available for all modelers. Ten modelers produced candidate models, with each modeler responsible for a subset of the data splits. However, each modeler was required to create candidate models using at least two of the algorithms available for use in Clementine: decision trees, rule induction, neural networks, and association rules. Over 90 models were generated at the end of the modeling phase. During this stage the case study included several of the developed models in the final system in order to reduce the variance and bias of the final selection of results on the large unseen data set.

Because of multiple modelers, a standardized method of documentation was established so that all models could be evaluated equally. Within Clementine, models were set up to produce the total number of:

1. Payments labeled as fraudulent and correctly called fraud;
2. Payments labeled as Non-fraudulent and correctly called non-fraud;
3. Payments labeled as fraudulent and incorrectly called non-fraud;
4. Payments labeled as Non-fraudulent and incorrectly called fraud.

In order to validate the cast study strategy and methodology, they included the eleven best models in the final decision-making process. The goal was to minimize the false alarms and maximize sensitivity. The final decision was based on the majority vote; if most of the final models classified a payment as "suspicious," that payment was labeled an "anomaly" and a candidate for further investigation. With 11 total models, a majority vote meant that six or more models must flag the 7 payment.

**Summary and Conclusions**

This case study has been outlined for overcoming problems common in fraud detection applications, including paucity of known fraud data, large amounts of unlabeled data which is nearly always non-fraudulent, and using as much of the data as possible so that all of the known patterns of fraud are captured. The procedure required paying careful attention to subtle database issues, such as the dependence of payments in the data, before splitting the data into training, testing, and validation data sets.

The results also shows the benefit of model ensembles to reduce the risks associated with the selection of a single model, to improve model sensitivity and to reduce false alarms. It was unclear for the author which of the factors in the data design and modeling stages contributed the most to the final favorable results, or how many models were necessary to obtain the benefits of combining model outputs. The authors answers to questions have practical significance and is explored in future work.

# Works Cited

Abbott, D. (2014). *Applied Predictive Analytics.* Indianapolis, Indiana: John Wiley & Sons, Inc.

Dean Abbott, Vafaie, H., Hutchins, M., & Riney, D. (2000). *Improper Payment Detection in Department of Defense.* Washington, DC: Federal Data Mining Symposium.