Six Sigma Methodology with Fraud Detection

ANDREJ TRNKA
Department of Applied Informatics
University of SS. Cyril and Methodius in Trnava
Nam. J. Herdu 2, 917 01 Trnava
SLOVAKIA
andrej.trnka@ucm.sk    http://fpv.ucm.sk

Abstract: - This paper describes one of the selected Data Mining task. It deals with implementation of Data Mining algorithms to fraud detection. The first part briefly introduces the applications of Data Mining and then is given to the fraud detection. Case study of fraud detection is focused to the field of application of financial grants for farms. It uses fictitious data.

Key-Words: - Six Sigma, Data Mining, Fraud Detection

1 Applications of Data Mining
The aim of data mining is to make sense of large amounts of mostly unsupervised data, in some domain. Systematic exploration through classical statistical methods is still the basis of data mining. Some of the tools developed by the field of statistical analysis are harnessed through automatic control (with some key human guidance) in dealing with data. [5], [6] Fraud detection [1] is relevant field of application of data mining. Fraud may affect different industries such as telephony, insurance (false claims) and banking (illegal use of credit cards and bank checks; illegal monetary transactions). In many situations it is the detection of outliers in the data that is most interesting. For example [2], the detection of fraudulent insurance claim applications can be based on the analysis of unusual activity. Hawkins [4] defines an outlier as an observation which deviates so much from other observations so as to arouse suspicions that it was generated by a different mechanism.

There is a rapidly growing body of successful applications [3] in a wide range of areas as diverse as:
- analysis of organic compounds,
- automatic abstracting,
- credit card fraud detection,
- electric load prediction,
- financial forecasting,
- medical diagnosis,
- predicting share of television audiences,
- product design,
- real estate valuation,
- targeted marketing,
- thermal power plant optimization,
- toxic hazard analysis,
- weather forecasting…

2 Fraud Case Study
This demonstration example [8] shows the use of Data Mining algorithms and methods in detecting behavior that might indicate fraud. The domain concerns applications for agricultural development grants. Two grant types are considered:
- arable development,
- decommissioning of land.

The example uses fictitious data to demonstrate how analytical methods can be used to discover deviations from the norm, highlighting records that are abnormal and worthy of further investigation. The real data can be storage in data warehouses. [9] The applicant is particularly interested in grant applications that appear to claim too much (or too little) money for the type and size of farm. The data set contains nine fields:
- id. A unique identification number.
• name. Name of the claimant.
• region. Geographic location (midlands/north/southwest/southeast).
• landquality. Integer scale – farmer’s declaration of land quality.
• rainfall. Integer scale – annual rainfall over farm.
• farmincome. Real range – declared annual income of farm.
• maincrop. Primary crop (maize/wheat/potatoes/rapeseed).
• claimtype. Type of grant applied for (decommission_land/arable_dev).
• claimvalue. Real range – the value of the grant applied for.

For building the model we used IBM SPSS Modeler. To do a first screening for unusual records, we used the anomaly detection. After identifying the input variables and executing, the Anomaly detection model was generated. Figure 1 shows the results with potential anomalies. The overall anomaly index value is also listed for each record, along with the peer group and the three fields most responsible for causing that record to be anomalous.

For building the model we used IBM SPSS Modeler. To do a first screening for unusual records, we used the anomaly detection. After identifying the input variables and executing, the Anomaly detection model was generated. Figure 1 shows the results with potential anomalies. The overall anomaly index value is also listed for each record, along with the peer group and the three fields most responsible for causing that record to be anomalous.

This helps to form hypotheses that can be useful in modeling. We used charts to gain a better picture of which records are being flagged. However, to understand relationships, it is worth taking a closer look at the data. Data investigation helps to form hypotheses that can be useful in modeling. Initially, we consider the possible types of fraud in the data. One such possibility was multiple grant aid applications from a single farm. Figure 2 shows multiple claims.

Based on this, we discarded records for those farms that made multiple records. Next step was to focus on the characteristics of a single farm applying for aid. We built a model for estimating what we would expected a farm’s income, based on its size, main crop type, soil type and so on. To prepare for modeling, we needed to derive new fields (f.e.: farmincome * rainfall * landquality). To investigate those farmers who deviated from the estimate, we derived another field that compares the two values and returns a percentage difference: 

\[
\frac{(\text{abs}(\text{farmincome} - \text{estincome}) \times 100)}{\text{farmincome}}
\]

(1) Figure 3 shows the histogram of this new field.
Since all of the large deviations seem to occur for arable_dev grants, it made sense to select only arable_dev grant applications.

From the initial data exploration, it was useful to compare the actual value of claims with the value one might expect, given a variety of factors. Using the variables in data set, the neural net or other methods we made the predictions based on the target, or dependent variable. Using these predictions, we explored records or groups of records that deviate. Figure 4 shows comparison predicted (by neural net) and actual claim values. It appears to be good for the majority of cases.

By deriving ((abs(claimvalue - 'S\text{-}N\text{-}claimvalue') / 'claimvalue') * 100) (2), the new field was created. It is similar to the field derived earlier. In order to interpret the difference between actual and estimated claim values, we used a histogram of this new field. We were primarily interested in those who appear to be claiming more than we would expect. By adding a band to the histogram, we selected records with a relatively large value, such as greater than 50%. These claims warrant further investigation.

3 Results
The result is the table with large value of new field. Figure 6 shows the potential fraudsters.
This paper demonstrated two approaches for fraud detection – anomaly detection and a modeling approach based on a neural net. In our real research we try to implement fraud detection to Six Sigma methodology. Figure 7 shows location of fraud detection in Control phase of Six Sigma methodology.

**4 Acknowledgment**

Grateful acknowledgment for translating the English edition goes to Juraj Mistina.

This paper was supported by Institutional grant of University of SS. Cyril and Methodius in Trnava.

**References:**


