**Paper 345-2010**

# Fraud Detection with SAS® Data Mining

Sascha Schubert, SAS Institute Inc., Cary, NC

## ABSTRACT

Fraud is a significant problem in many industries, such as banking, insurance, telecommunication, and public service. Detecting and preventing fraud is difficult, because fraudsters develop new schemes all the time, and the schemes grow more and more sophisticated to elude easy detection. Many organizations have implemented fraud detection and prevention systems based on SAS data mining to help them stay ahead of the fraudsters and avoid losing money.

This presentation will provide an overview of different data mining techniques that have proven successful in detecting different types of fraud. Using case studies, successful implementations in different industries will be described.

## INTRODUCTION

Organizations are exposed to fraud in many different ways. Fraud can be described as a group of intentional acts made for personal gain by causing significant losses to organizations. Besides the losses, their reputation can be affected, as well as their customers' loyalty.

Using insurance as an example, Figure 1 shows that there is a wide spectrum of fraud, ranging from opportunistic to premeditative. *Opportunistic fraud* is committed when the opportunity arises without much planning going into it (for example, when a car accident occurs, the claimant decides to inflate the claim to receive more money than he or she is entitled to). Usually, with this type of fraud, only one person or a very small group of people is involved. On the other side of the spectrum, there are potentially large groups of offenders who invent schemes to defraud insurance organizations. For example, they stage accidents to make large — and illegal — claims for fake injuries and car damage against an auto insurance company.

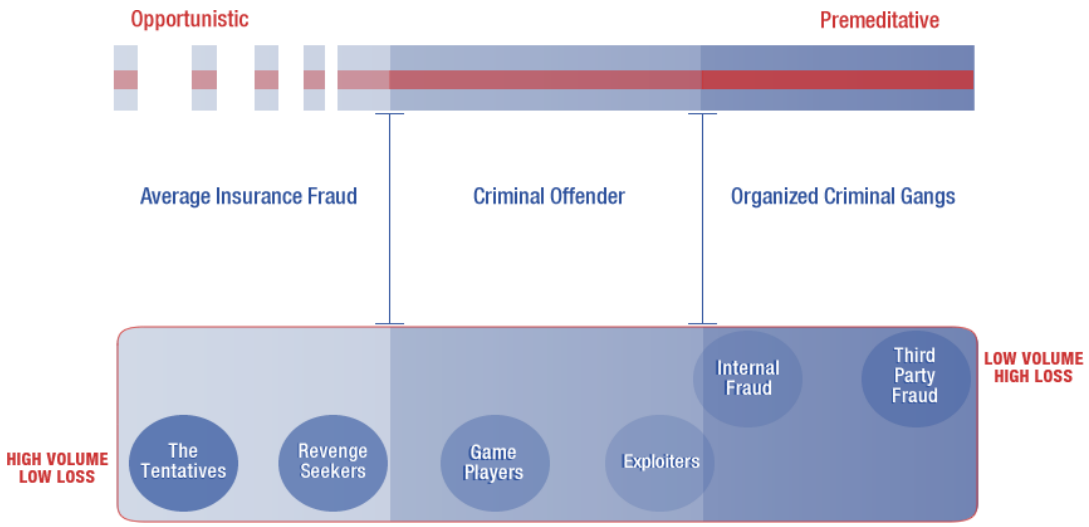


**Figure 1: Spectrum of Fraudster Types in Insurance**

Fraud is a very apparent problem across all industries. It ranges from external fraud committed by offenders outside the organization, either customers or not, to internal fraud, where employees of the organization are part of the scheme.

A proactive approach to fraud detection helps to reduce the effect of fraud on organizations. Through the use of advanced analytics, such as data mining, organizations are able to detect ongoing fraud earlier and take preventive measures to minimize losses.

## ANALYTICAL APPROACHES TO FRAUD DETECTION

The success of SAS in helping organizations to detect and manage fraud in a more proactive way is based on the integration of many different approaches to fraud detection. In Figure 2, this hybrid approach is illustrated. Historically, fraud detection systems have relied primarily on business rules. This has been good for identifying reoccurrences of lessons that have already been learned. With a hybrid approach, the organization extends its arsenal of fraud detection tools and starts looking for unknown and complex patterns that they were not able to see before.

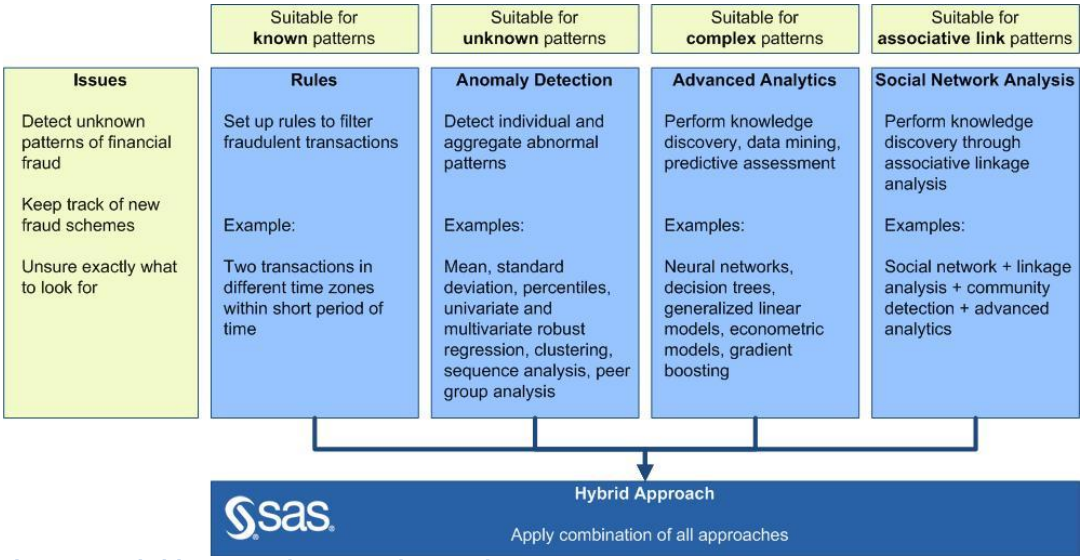


**Figure 2: Hybrid Approach to Fraud Detection**

In this paper, we will focus on the application of data mining techniques to detect anomalies and to build predictive models.

## DETECTING ANOMALIES

Techniques in this group are usually applied to fraud detection when the fraud pattern is not known and, thus, no target is available to use predictive modeling techniques. There are a wide range of approaches available in the SAS arsenal, including analytical visualization for univariate and multivariate outlier detection, anomaly detection based on time series data, dynamic profiling, and cluster analysis for outlier detection, to name a few.

### ANALYTICAL VISUALIZATION FOR UNIVARIATE AND MULTIVARIATE OUTLIER DETECTION

Any visualization technique can be helpful to detect outliers. For example, a simple box plot on credit card transaction amounts or claims on auto insurance provide more insight than just looking at a table. Using statistical tests such as testing the observer distribution of the values against an expected distribution can help to easily and automatically detect outliers.

An interesting technique to test the distribution of monetary amounts that might be suspected to be fabricated by a fraudster is Benford's Law. Based on the plausible assumption that people who make up figures tend to distribute their digits fairly uniformly, a simple comparison of the first-*n*-digits frequency distribution from the data with the expected distribution according to Benford's law is likely to reveal any anomalous results (see the Wikipedia entry list in the References section).

Benford's law states that the probability of encountering a number starting with the string of digits *n* is given by the following:

$$\log_{10}(n+1) - \log_{10}(n) = \log_{10}\left(1 + \frac{1}{n}\right).$$

An insurance company applied this technique to monitor claim amounts in order to detect amounts that are suspicious. They found that they had a suspiciously high frequency of amounts starting with the two digits 24 (see Figure 3). It was revealed that these were amounts just below Euro 250 because this was the threshold for claims that would be paid through the fast track without detailed checking. Further analysis also revealed that many of the claims with these amounts were filed for windscreen damage. The investigation revealed that the claims were false claims that had gone unnoticed.

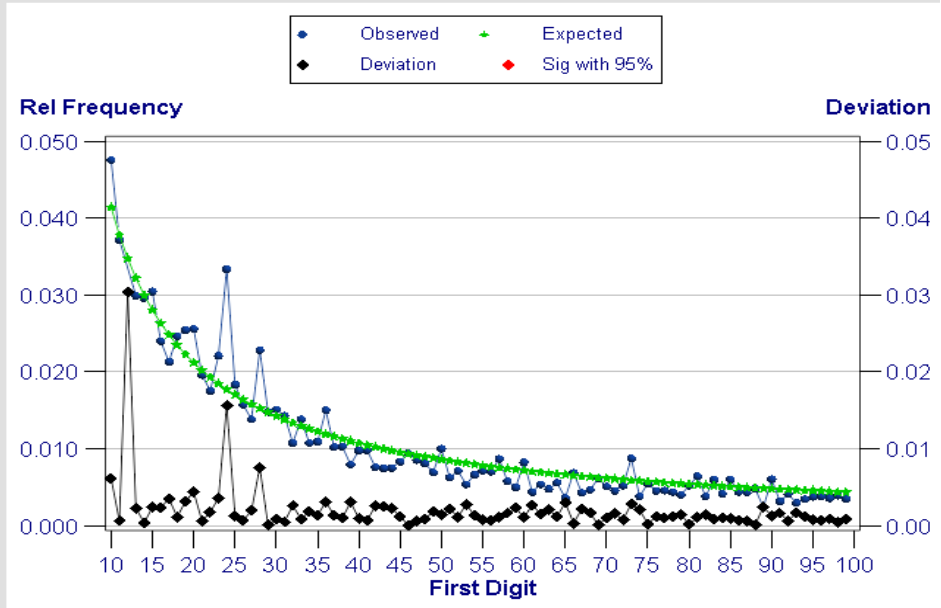


**Figure 3: Observed vs Expected First-2-Digits Distribution of Claim Amounts**

## ANOMALY DETECTION BASED ON TIME SERIES DATA

Figure 4 shows the forecast for the average time to report an accident using auto insurance data. We can see that most of the data points follow the predicted pattern; however, one point shows a substantial deviation. With the use of time series analysis techniques, analysts are able to take seasonal patterns into account for identifying deviations from the norm. In many cases, organizations might have business rules for detection of abnormalities on predefined metrics, such as the time to report an accident. If the threshold implemented with the business rule is exceeded, the rule would create an alarm in the process. However, these rules tend to be rather static and do not take observed fluctuations into account.

Regular time series analysis combined with forecasting can automatically take these natural fluctuations into account. For example, during summer vacation time, it might take longer to report accidents as can be seen in the data in Figure 4. Also, there seems to be a long-term trend that average reporting time has increased. Statistically significant patterns are detected by the time series analysis and can be taken into account for the creation of seasonally adjusted thresholds for the average reporting time.

This enables organizations to be much more dynamic in their fraud detection approach.

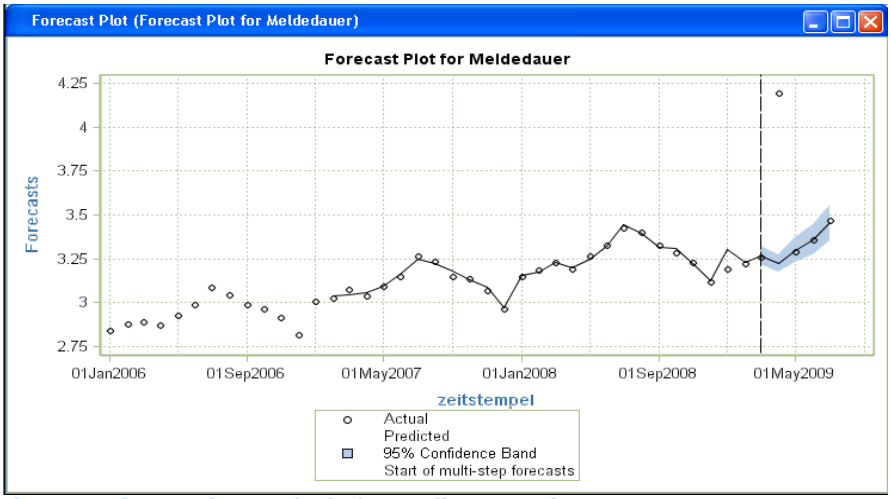


**Figure 4: Time Series Analysis for Outlier Detection**

## DYNAMIC PROFILING

For time series data, dynamic profiling can be applied to create automatic alerts based on unexpected changes in the behavior profile (see Figure 5). For each dimension to be monitored, metrics are calculated based on historical data. For example, for credit card usage, each customer's average monthly number of transactions, monthly balance, and monthly revenue are calculated. Using variability statistics, thresholds are created for the monthly expected behavior for these measures. Here's an example:

Expected maximum daily revenue = 3 months average maximum daily revenue ± 3 times standard deviation.

The actual behavior is then constantly monitored against this profile and alerts are created when thresholds are exceeded. The profile is updated monthly to capture longer term changes in customer behavior. If historical data is not available, peer group analysis can provide the profile statistics. For example, a new credit card customer can be assigned to a group of peers based on demographics, product portfolio, and usage intentions.

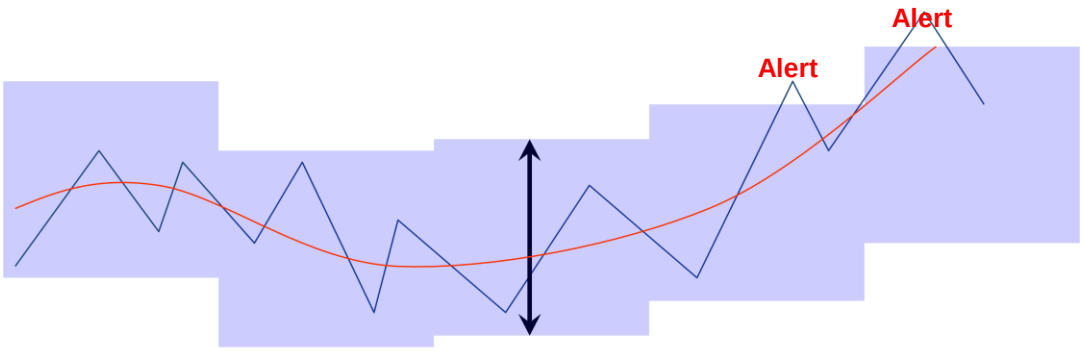


**Figure 5: Dynamic Account Behavior Profiling**

## CLUSTER ANALYSIS FOR OUTLIER DETECTION

Cluster analysis has been the technique of choice for multivariate outlier detection. Figure 6 shows an example of the cluster distance plot created on insurance agent claims data. The suspicion here is that some agents are involved in fraud cases by colluding with customers and repair shops. We can see that three clusters (Cluster 4, Cluster 8, and Cluster 9) are located at the edge of the cluster cloud, which potentially points to outliers captured in those clusters.

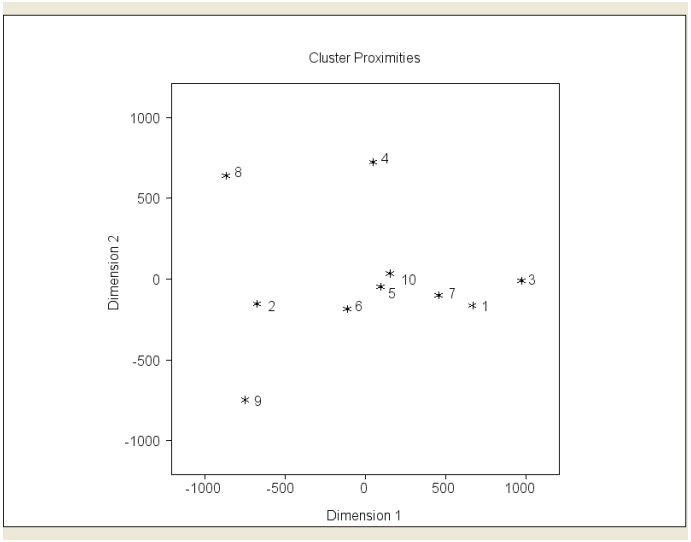


**Figure 6: Cluster Distance Plot**

We can use the Segment Profiler in SAS© Enterprise Miner™ to get a profile of these abnormal clusters (see Figure 7). When overlaying confirmed fraud cases with the results of the cluster profiling, it shows in Figure 7 that Cluster 8 captures a large portion of the known fraud. Note that the flag of known fraud was not used for the segmentation. It was used only later for reporting on the cluster characteristics. The very high percentage of male customers in the customer portfolio of an agent, as well as the low number of repair shops the agent is cooperating with, point to a high risk of fraud. Using these characteristics, an insurance company can now set up high-risk profiles of agents and monitor their agent portfolio against these rules.

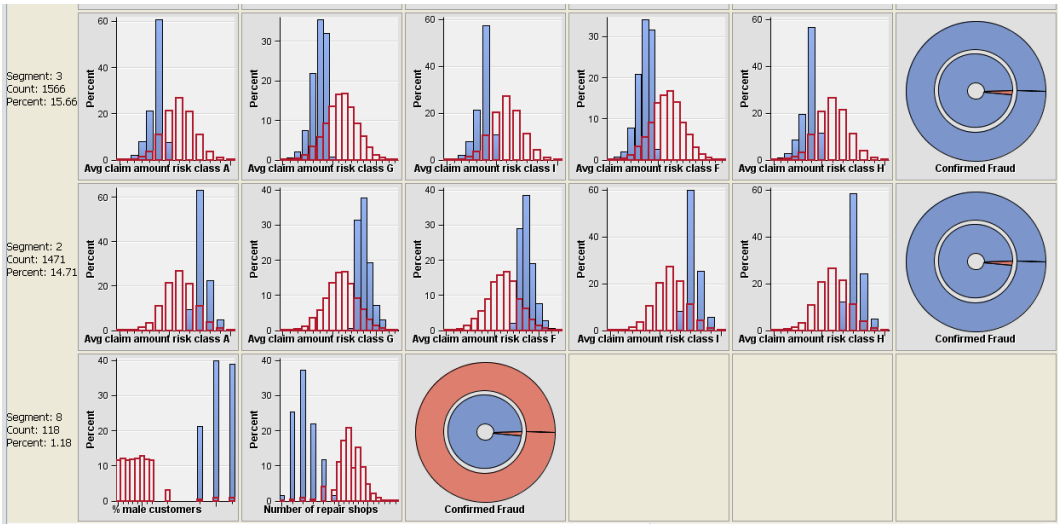


**Figure 7: Cluster Profiles**

## OUTLIER DETECTION BASED ON PREDICTIVE MODELING TECHNIQUES.

If no target for confirmed fraud is available from the data, the data attribute under investigation can be modeled using predictive modeling techniques and outliers can be detected using the residuals. For example, let's assume we want o analyze amounts from doctors' claims in health care. Let's make the claim amount itself the target for a predictive model and use all the other data that is available around the claim as input to the predictive model.

The model now predicts an expected claim amount using the relationships between the input and the target. If the model performs reasonably well, an assumption that we can check using our favorite model performance measurement, we can conclude that amounts that are not predicted well and create large residuals between predictions and observed values are cases that might be suspicious.

We can now easily define a measurement that identifies the extreme residuals, such as the three standard deviations percentile and mark all residuals that exceed this threshold for further investigation.

For example, Figure 8 shows the analysis of the residuals between predicted and observed doctors' billing amounts. The graph suggests that amounts larger than 45 units are rather unusual and should be considered abnormal. This graph was created using the residual output of a regression model and running the UNIVARIATE procedure on the output.

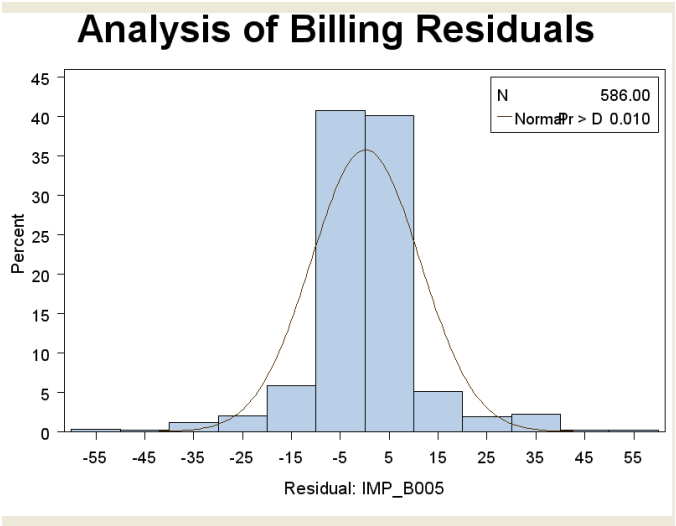


**Figure 8: Distribution of Residuals from Billing Data Predictions**

There are certainly other analytical techniques that can help to detect outliers in multidimensional data. All these analytical techniques have the advantage of providing thresholds for outliers based on dynamic data analysis rather than on instinct. Furthermore, they can take multivariate relationships between different dimensions into account and they are easy to keep up-to-date through recalibration of the underlying model.

## LINK AND NETWORK ANALYSIS

In contrast to opportunistic fraud, organized crime often involves a group of people who work together. To uncover these groups, organizations need to be able to extend the analysis beyond single accounts. Techniques like link analysis, association analysis, and path analysis can help to uncover relationships that point to suspicious behavior.

Figure 9 illustrates the outcome of the application of path analysis to stock trading data. Here we have a case of ring trading on the stock market. By applying path analysis using the ID of stock certificates, it is possible to visualize and analyze the relationship between sellers and buyers. In this case, we see that a bond is sold between different dealers before it ends up with the same dealer in the end. This is behavior that might be investigated further.

| Rule |
| --- |
| Dealer 7 ==> Investor Group 2 ==> Dealer 7 |
| Dealer 7 ==> Investor Group 2 ==> Dealer 7 ==> Investor Group 2 ==> Dealer 7 |
| Dealer 7 ==> Investor Group 2 ==> Dealer 7 ==> Investor Group 2 ==> Dealer 7 ==> Investor Group 2 ==> Dealer 7 |
| Dealer 7 ==> Dealer 17 ==> Dealer 7 |
| Dealer 7 ==> Investor Group 2 ==> Dealer 7 ==> Dealer 17 ==> Dealer 7 |
| Dealer 7 ==> Investor Group 2 ==> Dealer 7 ==> Investor Group 2 ==> Dealer 7 ==> Dealer 17 ==> Dealer 7 |
| Dealer 7 ==> Investor Group 2 ==> Dealer 7 ==> Investor Group 2 ==> Dealer 7 ==> Investor Group 2 ==> Dealer 7 ==> Dealer 17 ==> Dealer 7 |
| Dealer 17 ==> Dealer 7 ==> Dealer 17 |
| Dealer 7 ==> Dealer 17 ==> Dealer 7 ==> Dealer 17 ==> Dealer 7 |
| Dealer 7 ==> Investor Group 2 ==> Dealer 7 ==> Investor Group 2 ==> Dealer 7 ==> Dealer 17 ==> Dealer 7 ==> Dealer 17 ==> Dealer 7 |
| Dealer 7 ==> Investor Group 2 ==> Dealer 7 ==> Dealer 17 ==> Dealer 7 ==> Dealer 17 ==> Dealer 7 |
| Dealer 7 ==> Dealer 17 ==> Dealer 7 ==> Investor Group 2 ==> Dealer 7 |

**Figure 9: Ring Links in Trading Data**

SAS now offers SAS® Social Network Analysis, which enables organizations to monitor customer behavior across multiple accounts and systems in order to analyze all related activities and relationships at a network dimension. For more information about SAS Social Network Analysis, please refer to this Web page: http://www.sas.com/solutions/fraud/social-network.

## SUPERVISED MODELS

Predictive modeling is by far the most common analytical technique applied to fraud detection when the fraud is known and historical data exists. Using supervised modeling such as generalized linear and non-linear models, decision trees, and neural networks, a fraud risk score is calculated to identify suspicious behavior similar to previous fraud patterns. Besides quantifying probability of fraud for each record, the model can help to reveal and describe the fraud patterns in more detail. However, with fraud, it is often more important to create a model that predicts as accurately as possible in order to minimize the number of false alarms. For this reason, rather complex neural networks are quite common.

False alarms lead to wasted resources in terms of investigation time and personnel. False accusations of customers can also have a significant impact on brand reputation and customer satisfaction. Because it is often unknown which modeling technique will provide the best performance, testing different approaches against each other in an efficient manner is critical. SAS Enterprise Miner provides a large arsenal of predictive modeling algorithms in an intuitive and scalable workflow environment.

The performance of the predictive model can be assessed using an ROC chart, as shown in Figure 10. The information gained from this chart helps compare individual models against each other as well as against random decisions. The larger the area between the diagonal baseline, which depicts the random decision and the curve depicting the model, the better the predictive performance of the model is. In order to provide an unbiased performance evaluation, the ROC chart is also created on unseen validation data and compared against the training data that was used to calibrate the model. So, in short, the farther to the left upper corner of the chart the model curve is located, the better the model performance.

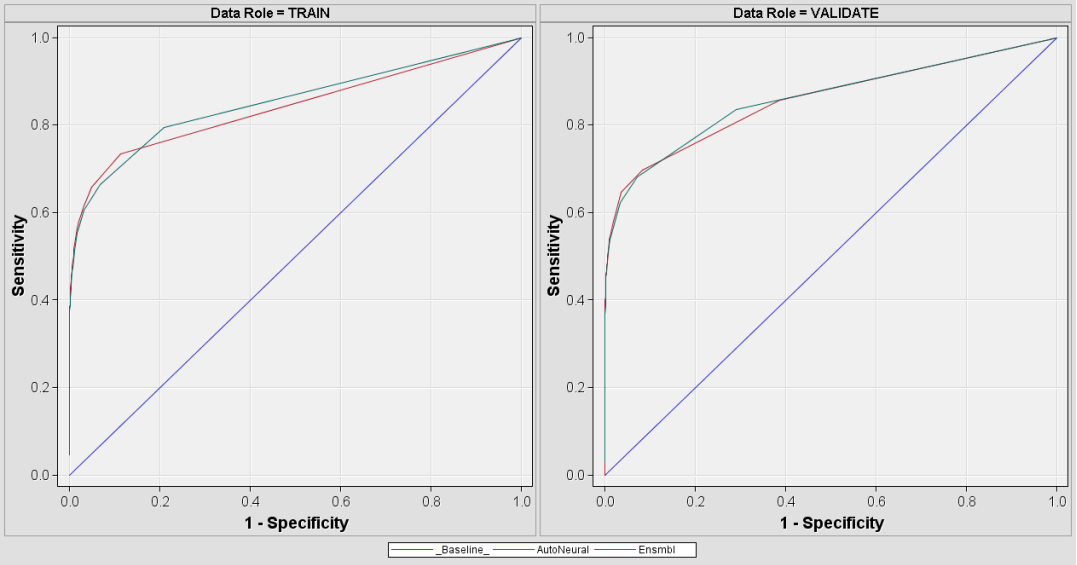


**Figure 10: ROC Chart for Predictive Model Performance Evaluation**

## RARE TARGETS

A common obstacle for predictive fraud detection models is the rare occurrence of the target event. The proportion of identified fraud cases in the model input data is usually very small, which can make it difficult to extract the fraud signal in the data.

SAS provides several techniques that help overcome the issue of rare targets.

### OVERSAMPLING WITH SUBSEQUENT DECISION ADJUSTMENT

In order to avoid predictive modeling that ignores rare targets, the target event proportion is adjusted using oversampling techniques. Often, oversampling is applied to create samples with equally sized target event proportions, where both event values have a proportion of 50%. In the few fraud cases in the historical data, this method has the disadvantage of excluding the overwhelming majority of the non-event records. For example, in a data set of 10,000 records with 100 fraud cases, equal size sampling would create a data set with all 100 fraud cases and 100 randomly selected non-fraud cases. This massive shrinking of the data set might cause information loss. That is why other oversampling event distributions should be considered, such as 80%of non-cases and 20% of cases.

SAS Enterprise Miner provides a wide range of oversampling methods for the user to easily define the desired event distribution. In concert with oversampling, the predictive models should be adjusted for the biased prior probabilities. SAS Enterprise Miner features comprehensive decision processing that enables the user to integrate the decision adjustments into the data mining process flow (see Figure 11).

### RULE INDUCTION

SAS Enterprise Miner also features an integrated predictive modeling technique to automatically deal with rare events in input data sources. This modeling algorithm successfully searches for pure sections in the database that contain either none or very few of the event cases. This segment is then removed from the input data, and the resulting data is searched again until the optimal segmentation is found in relation to the target.

This algorithm does not require any prior oversampling or post-modeling decision adjustments and simplifies the predictive modeling for rare events. For more information about the Rule Induction model node, please refer to the SAS Enterprise Miner 6.1 online Help.

As can be seen in Figure 11, the Rule Induction node helps to minimize the number of process steps that a user has to build to properly deal with rare events in input data for predictive models. When using a traditional oversampling approach, use the following nodes to achieve the desired goal:

- Input Data source node to access the input data

- Sample node to oversample to target event records (confirmed fraud cases)
- Data Partition node to partition the data into training, validation, and testing data
- Model nodes (Decision Tree and Regression) to train a predictive model
- Decision node to adjust the model decisions back to original rare event proportions (priors)

When you are using the Rule Induction node, the same goal can be achieved in only three steps because no oversampling and decision adjustment is required.

- Input Data source node to access the input data
- Data Partition node to partition the data into training, validation, and testing data
- Model nodes (Decision Tree and Regression) to train a predictive model

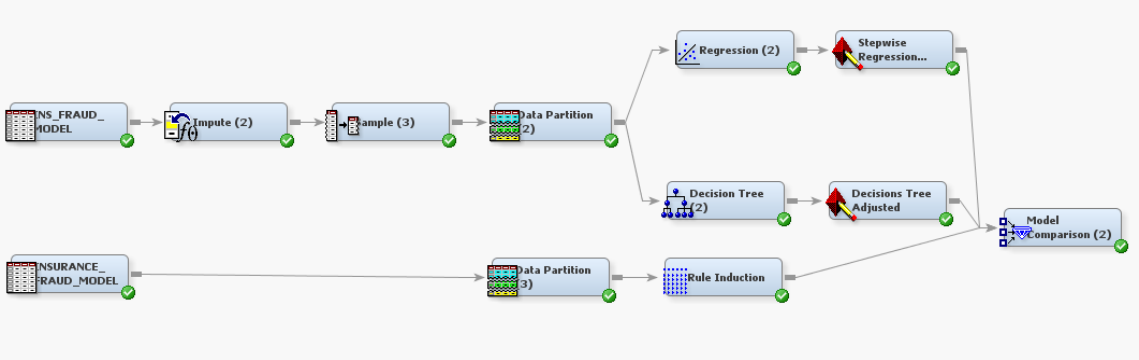


**Figure 11: Process Flow for Rule Induction Model in SAS® Enterprise Miner™**

## SAS SOLUTIONS FOR ENTERPRISE FRAUD MANAGEMENT

Data mining and predictive modeling are crucial parts of an enterprisewide fraud detection and prevention strategy. However, for optimal fraud management, the models need to be integrated into an enterprise-wide process. The SAS© Fraud Framework provides a technology infrastructure that integrates fraud detection, alert management, network analysis, and case management — giving organizations the upper hand in detecting fraud in any form at any touch point. This framework meets the needs of multiple lines of businesses, channels, and industries. This framework is flexible, leveraging different SAS technologies, data management capabilities, and analytical techniques depending on the type of fraud and the industry. It is available for different industries: banking, insurance, heath care, and government. For more information about the SAS Fraud Framework, please refer to this site: http://smportal.sas.com/products/fraud/Pages/default.aspx.

## CONCLUSION

Fraud is a major threat across all industries. With the evolution of technology, fraud techniques have improved and are more undetectable than ever. Using analytics, organizations can protect themselves better against losses from any kind of fraud. The variety and flexibility of data mining techniques offered by SAS helps to extract the right insight from data that enables organizations to implement the optimal strategy against fraud.

### REFERENCES

Benford's law, Wikipedia: http://en.wikipedia.org/wiki/Benford's_law
SAS Data Mining: http://smportal.sas.com/products/analytics/data-mining/Pages/default.aspx
SAS Social Network Analysis: http://www.sas.com/solutions/fraud/social-network
SAS Fraud Framework: http://smportal.sas.com/products/fraud/Pages/default.aspx

### ACKNOWLEDGMENTS

The author would like to thank the following SAS employees for their valuable contributions to this paper: Stefan Ahrens, SAS Germany, and Colin Gray and David Hartley, SAS United Kingdom.

### CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Sascha Schubert

SAS Institute Inc.
Domaine de Grégy
Grégy-sur-Yerres
77257 Brie Comte Robert Cedex
E-mail: Sascha.Schubert@sas.com