USPS Worker's Compensation Fraud Detection Model
A Statistical Approach

Background Provided: To contain the escalating medical and other personnel expenses related to the benefits paid by the Office of Workers' Compensation Programs (OWCP), the US Postal Service's Risk Analysis Research Center, wishes to develop a predictive fraud model. Such a model will help to identify high risk claims, score claims based on their likelihood of fraud, magnitude of such claims and the key drivers (independent variable) for such activity. To achieve its objectives, the OWCP has outlined a general two-phase approach and has requested a contractor to provide further input to the analytic approach. The first phase, Data Acquisition and Feasibility Assessment, will entail data access, data processing and organization of data for model building. The second phase, Development of an Analytic Tool, will entail the actual data analysis and model building in consultation with the USPS-OIG.

Approach: The contractor understands that insurance fraud costs associated with workers' compensation are a major portion of the $25B property and casualty loss in the country. Their own estimates range up to $8B in terms of future liability. A portion on this liability is caused by fraudulent activity or false claims. Our initial approach would be to define fraud into these three or more possible categories:

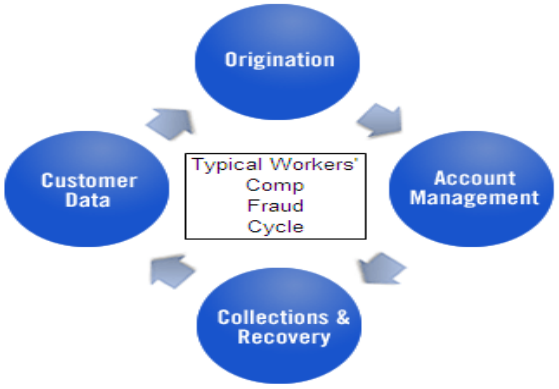
1. Health Insurance Practitioner Fraud: Where the provider/practitioner issues bills for services not rendered or renders services where there is no injury.
2. Coverage Fraud: Where claimant files claims for injury/pain where there is none.
3. Agency Fraud: Where employees within the OWCP generate fictitious claims for those covered to a partner in crime.

Once a classification is developed, the magnitude of fraud in each of these categories will be assessed based on prior experience of claim auditors or supervisors.

Period of Interest: For the initial study, a time period will be established that would give us the highest probability of identifying fraudulent activity data availability. Data may be coming from various sources for legacy data whereas more recent data may be readily available.

Domain Expertise: In depth interview will be conducted by the contractor to understand the nature of fraudulent cases from supervisors or claims auditors. They can provide very useful information when creating derived variables for analysis. For instance, if fraudulent calms come from providers that are not located in the same geographic region as the place of injury, a derived distance variable would have to be created in the model to score high risk cases.

Process Cycle Example:



Data Access and Variables of Interest: During the initial feasibility assessment phase, information will be gathered on the databases available, their OS, size, attributes, storage issues, reliability, etc. The key fields of interest will be explored and their availability for analysis. As indicated in the project summary, the key variables of interest will include the following:

- Gender
- Age (at time of injury, current age, etc.)
- Type of Injury
- Number of Prior Claims
- Schedule Award Payment
- Third Party Claim
- Occupational Code
- Amount of Medical Billing
- Amount of Prescription Billing
- Employment Location
- PO Box as Mailing Address
- Someone at Same Residence with Workers Compensation Claim
- Geographic Location of Injury
- Relocated Employee
- Injury from Hire Date
- Stress Claims
- Provider Location and Zip code
- Risk Factors for Disease
- Provider Geographic Location Vs. Injury Location
- ICD9, CPT4 Codes

Summary Statistics: Data will be analyzed for reliability, validity and completeness. Further, data will be classified into multiple sub-sets to understand the relationship between various components to check accuracy from multiple sources.

Modeling Approach: We will use several approaches in the model development phase. Clustering and variable selection methods will be used to identify available variables or derived variables that are predictive of the outcomes. We will then develop flexible and robust regression strategies to model (a) fraud and (b) the amount of fraud, providing thresholds that will raise flags for further investigation.

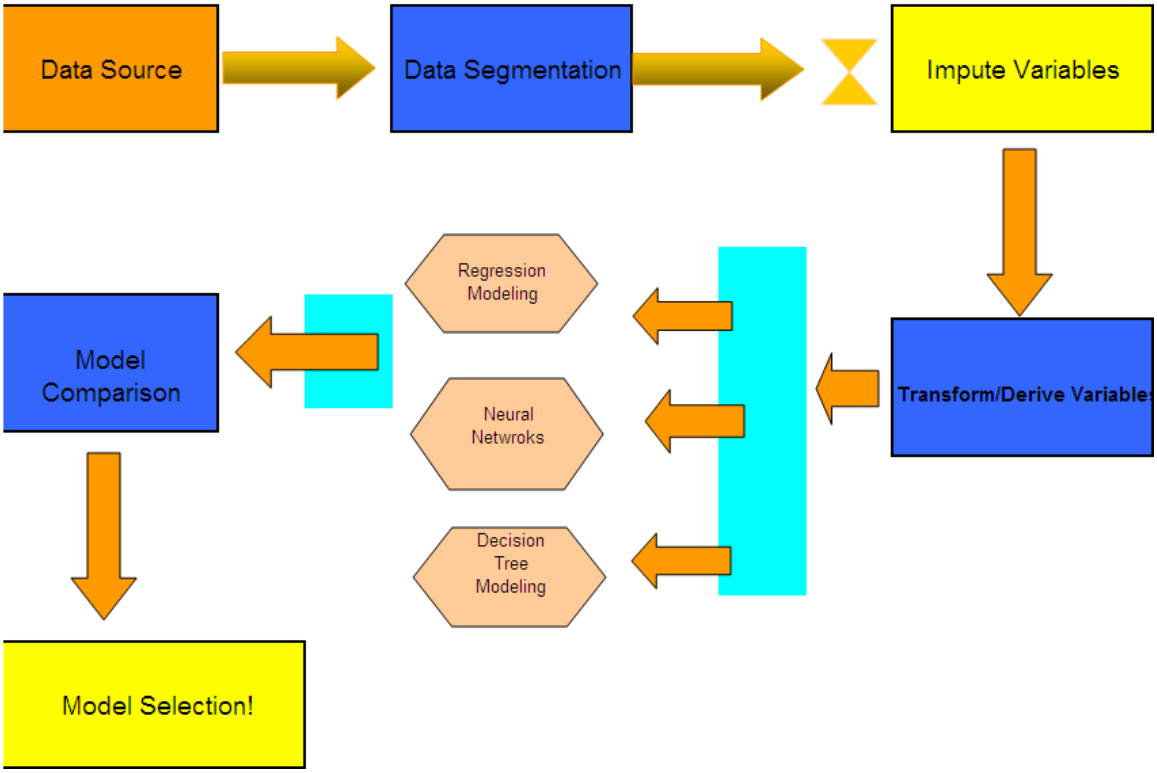
1. **Data and Variable Segmentation:** We will initially use clustering methods like hierarchical clustering to segment both the variable space and the sample space into homogeneous clusters. The role of clustering is two-fold. First we will cluster the variables of interest to determine which variables act similarly and thus would be redundant in the model-building process. It is well known that the inclusion of several correlated variables in the model-building process is detrimental to model building and thus will be inefficient. It will also allow us to create "meta-variables" representative of the different kinds of variables present which will encapsulate the information in these variables. Secondly, we will cluster the cases to determine different case profiles that are present in the databases. We can potentially generate fraud detection and fraud magnitude models within each class of cases, which will reduce our prediction error. The clustering score code can then be used to score all cases based on which cluster they belong to, and then those cases can be entered into the appropriate model for fraud prediction. This process thus segments our data into relatively homogeneous groups within which efficient models can be developed. Deliverable to the agency would be a method to obtain the cluster membership for each new case presented, thus allowing for the selection of an appropriate model. This will be implemented in basic SAS procedures that can be run at the client site.

2. Variable selection for Model building: We propose to use an ensemble of decision tree-based models (known as random forests to learn from the data which variables are most predictive of (a) fraud, and (b) high-dollar value fraud.[1] This is not part of the deliverables but an essential step needed to optimize modeling. Random forests are a very flexible and robust method of supervised learning which will allow us to efficiently select variables which are most predictive of fraud via the variable importance scores provided by the algorithm. The actual variables as well as derived variables (like different functions of variables, combinations of variables, "meta-variables" derived in

step 1) can all be incorporated in this feature selection process. This process will be carried out by the contractor with the aim of efficient feature selection that will enable us to develop a robust and accurate regression model for fraud

3. **Regression modeling for fraud detection and fraud value estimation**: The two steps outlined above are essential steps in developing appropriate models for this problem. Once we have selected the variables which would be most predictive of (a) the presence of fraud, and (b) the value of the fraudulent claim, the contractor will develop logistic regression models (for (a)) and linear regression models (for (b)) incorporating the variables (available and derived) that were identified in step 2. These models will then be further optimized using backward selection, and then their predictive performance assessed using 10-fold cross-validation within the available dataset. The goal will be to optimize the cross-validated prediction error to obtain the final deliverable models for each goal (a) and (b). We will also evaluate the predictive accuracy of (a) and (b) using receiver operating curves (ROC) and the area under the curves (AUC), which summarize the false positive and false negative rates at different thresholds. This will allow us to provide threshold values at which flags will be raised. The deliverables to the client will be logistic (a) and linear(b) models as well as thresholds based on the available database which are implemented in basic SAS procedures which can be run at the client site.

Modeling Process Flow: Let's assume data is available for Worker's Compensation from the USPS-OIG for the period 2000-2005. Let us also assume that included in the data are text fields added by the Adjuster or Supervisors. Search-string functions could be applied for detecting confirmed fraud cases. The flow diagram below shows a typical data extraction and modeling process:



Software Considerations: Based on the volume of data, the operating system and the usability of the production systems, a choice will have to be made from a variety of modeling software available. The current contractor has experience with SAS, SPSS and Knowledge Seeker.

Conclusion: We propose a flexible and robust statistical modeling approach using clustering and ensemble methods to develop regression-based models for fraud detection and fraud value estimation. These models will be optimized for prediction error and thresholds developed using ROC methods to optimize false positive and false negative rates. Thus these models will be able to flag fraudulent activity as well as high-value fraud.