Fraud Detection: Data Mining Approach

Amir T Payandeh

Outline of talk

- What is Fraud
- Data mining and its process steps
- CASE STUDY 1: HealthCare insurance Fraud Detection
- Other Applications in Insurance Industry

Definition and types of fraud

We define fraud as any act designed to obtain an unlawful benefit from an insurance policy

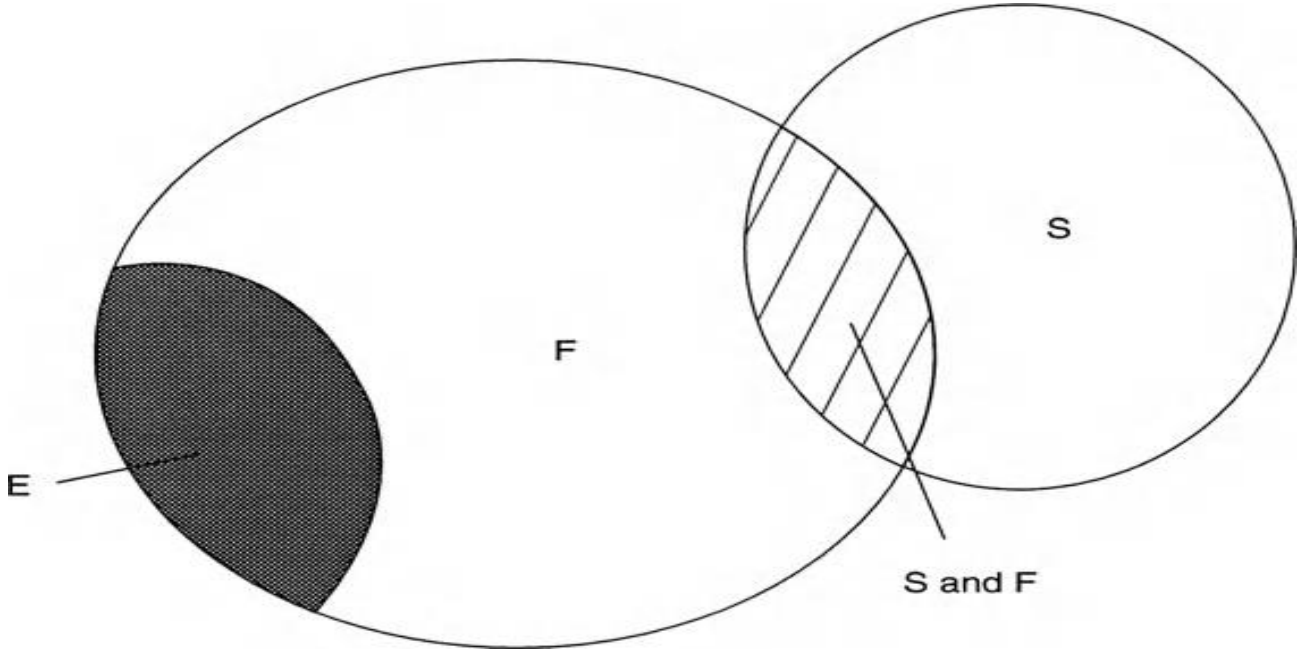
1. **Misrepresentation:** when someone gives false information in order to pay a lower premium
2. **Claim padding (build-up):** when someone increases the values of its claim to take financial advantage of the opportunity
3. **Claiming for non-existent loss (planned):** when someone declares a claim that did not occur to get rich
4. **Deliberately causing a loss:** when someone organizes a loss to make money from the insurance coverage

How important is

- 10 percent of auto insurance claims payments are attributable to fraud
- 3 to 6.4% of all claim payments contained fraud
- Fraud impacts on claiming patterns amongst which are and planned fraud
- fraudulent claims represent at least 13 percent of all claims
- 1/5 Americans believe it is okay to pad claims to make up for previously paid premiums or a policy's deductible

What is the problem with fraud estimation?

- The major problem when we have to evaluate the significance of fraud in a given market, is estimation.
- We cannot find easily a proportion of fraud over all coverage, because the numerator of this proportion is hidden information.
- The major statistical problem associated with these estimators is their lack of *robustness*



- Set F represents total fraud in the market
- Sets E and S show respectively the established and suspected fraud .

Data mining and its process steps

- Data mining combines data analysis techniques with high technology for use within a process.

- Its primary goal is to develop usable knowledge regarding future events.

The steps in the data mining process

1. Problem definition
2. Data collection and enhancement
3. Modeling strategies
4. Training, validation, and testing of models
5. Analyzing results
6. Modeling iterations
7. Implementing results

1-Problem Definition

- It is the most important step in the data mining process
- It should not be a discussion of the implementation or efficacy of techniques
- It should state the business objective in a clear and simple language
- It should state how the results are to be measured
- It should include estimates of the costs associated with making inaccurate predictions
- It should estimates of the advantages of making accurate

2-Data collection and enhancement

- Data mining algorithms are only as good as the Collected data
- Incomplete or biased data lead to incomplete or biased models with significant blind spots

Data collection steps

1. **Define Data Sources:** Select data from multiple databases
2. **Join and Denormalize Data:** Join the multiple data sources into a single file
3. structure
4. **Enrich Data:** fills gaps by adding more information to data (if necessary)
5. **Transform Data:** Examples: aggregating records, creating ratios, etc.

3. Modeling strategies

- Supervised learning methods are employed whenever there exists a target variable
- Unsupervised learning methods are employed whenever there does not exist a target variable

Modeling objectives

- Prediction: Prediction algorithms determine models or rules to predict target variable, given input data (predict the value of the S&P 500 Index)
- Classification: Classification algorithms determine models to predict discrete values given input data. (predict fraudulent Behavior)
- Exploration: Exploration uncovers dimensionality in input data (uncover groups of similar customers)
- Affinity: Affinity analysis determines which events are likely to occur in conjunction with one another (product purchase combinations)

Modeling Objective	Supervised	Unsupervised
Prediction	Regression and Logistic regression Neural Networks Decision Trees Note: Targets can be binary, interval, nominal, or ordinal.	Not feasible
Classification	Decision Trees Neural Networks Discriminant Analysis Note: Targets can be binary, nominal, or ordinal.	Clustering (K-means, etc) Neural Networks Self-Organizing Maps (Kohonen Networks)
Exploration	Decision Trees Note: Targets can be binary, nominal, or ordinal.	Principal Components Clustering (K-means, etc)
Affinity	Not applicable	Associations Sequences Factor Analysis

4. Training, validation, and testing of models

- Data partition into 3 data sets
  - 1th dataset to train a model
  - 2nd data set to validate the model
  - 3rd dataset to test the trained and validated model.
- This partitioning ensures us the model does not memorize a particular subset of data
- In some cases, controlling the splitting of training and validation data is desirable. For example, For instance, in fraud detection study we should control partitioning to avoid distributing the fraudulent events just into one partition.

5. Analyzing results

- **Linking Techniques to Business Problems**
- **Use diagnostic tools such as confusion matrix**. A confusion matrix involves a comparison of predicted values to actual values

Modeling Objective	Supervised (Target Field Data Exists)	Unsupervised (No Target Field Data Exists)
Prediction	Attrition/Retention Cash Used in ATM Cost of Hospital Stay Fraud Detection Campaign Analysis	Not feasible
Classification	Segmentation Brand Switching Charge Offs Fraud Detection Campaign Analysis	Segmentation Attrition/Retention
Exploration	Segmentation Attrition/Retention Scorecard Creation Fraud Detection Campaign Analysis	Segmentation Profiling
Affinity	Not applicable	Cross-sell/Up-sell Market Basket Analysis

CASE STUDY 1: HealthCare insurance Fraud Detection

- The first case study uses the data mining process to analyze instances of fraud in the public sector health care industry.
- In this study, the data contain recorded examples of known fraudulent cases.
- The objective of the health care case (business question) is to determine:
  - attributes depict fraudulent claims
  - patterns that produce fraudulent claims among program beneficiaries

- A public sector organization has collected data in which there exists a field indicating fraudulent activity.
- 15 input variable exist as inputs to predict fraud.
- The goals of the analysis are
  - to understand how the input factors relate to predicting fraud
  - to develop rules for identifying new cases to investigate

To chose a modeling strategy

1. **amount of missing data and how it is handled**
2. **measurement level of variables**: target field is binary; 14 inputs are nominal (with missing values) and 1 is interval variable (with no missing values).
3. **percentage of data representing the fraud**
4. **goal of the analysis**

Modelling Method

- Based upon facts 2 and 4, we can use the **regression, decision trees**, or **neural networks**
- training a neural network will be slow with the nominal inputs
- neural networks provide little information on how the inputs relate to the target event.
- Therefore, we **can't use neural networks** method
- About 7 percent of the records represent fraudulent cases
- many of the input variables have a high number of nominal levels. Therefore for regression method we need many observation.
- We **can't use regression** method

Possible results:

- 14 percent of the fraudulent records can be described as:
  - Male;
  - person categories=type A
  - received payments of between $19,567 and $44,500

- 9.8 percent of the fraudulent records can be described as:
  - Male
  - person categories=missing
  - payment status U
  - received payments of between $11,567 and $40,851

Other Applications in Insurance Industry

- **ADVERSE SELECTION**
- **Moral hazard**

- **In both cases: Typical approach is ”what is an optimal CHOICE of insurance contract?”**

Thanks