Deloitte.

Fraud in Insurance

Applications of Predictive Modeling

So What Exactly is Fraud?

Fraud is the crime of using dishonest methods to take something valuable from another person (definition of Fraud as given in Merriam Webster)

The legal definition of fraud defines it as means misappropriating assets or by deliberately misrepresenting or concealing material facts relevant to some financial decisions or by abusing responsibility, a position of trust (Section 17 of Indian Contract Act, 1872)

Insurance fraud occurs when any act is committed with the intent to fraudulently obtain some benefit or advantage to which they are not otherwise entitled or someone knowingly denies some benefit that is due and to which someone is entitled.

Trivia

According to an Indian association, Out of the total outgoings in health insurance, nearly 25% are fraudulent claims

Recent surveys conducted in the US show that more than 25% of respondents think it is acceptable to inflate insurance claims, even more believe it is reasonable to do so to recover deductibles

According to the Insurance Information Institute, the estimated fraud in US Property and Casualty industry is about $ 30 billion a year.

Deloitte.

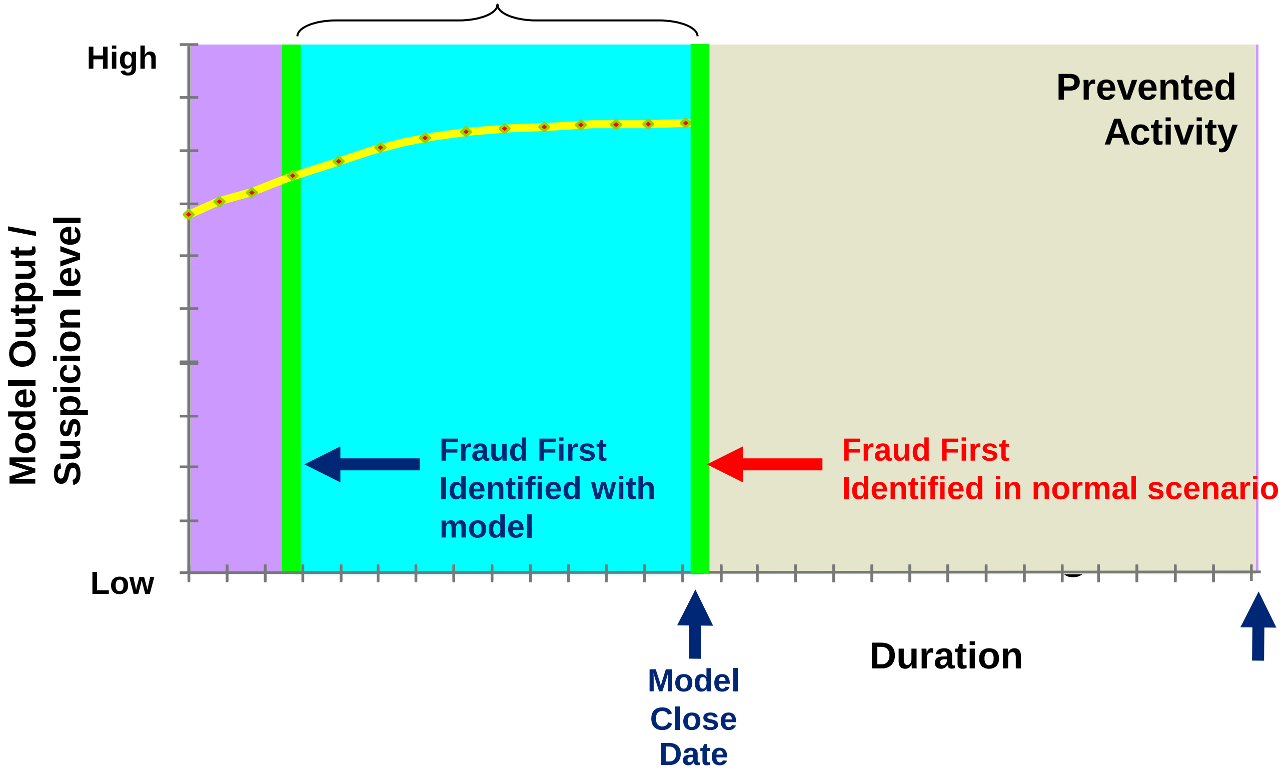
Types of Insurance Fraud

	Internal Fraud	Intermediary Fraud	Customer Fraud
Definition	Fraud against the insurer by its Director, Manager and/or any other officer, staff member	Fraud against the insurer or policy holders by an agent or any other third party administrator	Fraud against the insurer in the purchase or execution of an insurance product.
Examples	• Misappropriating funds • Fraudulent financial reporting • Forging signatures and stealing money from customers' accounts	• Non-disclosure or misrepresentation of risk to reduce premiums • Commission fraud – Insuring non-existent policy holders while paying premium to the insurer	**Soft Fraud:** • Exaggerating damages/loss • Deliberate or subtle lagging of claims resolution **Hard Fraud:** • Staging the occurrence of incidents • Medical claims fraud
Control Framework	Internal audit teams independently examine the processes and report weaknesses in control mechanisms	Having documented policy for appointment of new intermediaries, appropriate sanction policy in case of non-compliance by the intermediary	Adequate client acceptance policy, client should be identified and identity verified. Professional judgment based on experience should be used.

Deloitte.

Can Predictive Modeling be the Answer?

- Predictive modeling is the process of transforming data insights into an estimation of future outcomes upon which actionable decisions can be made
- With predictive modeling, one can identify fraud and refer the claim to fraud experts in less than 30 days, which under normal circumstances could take 3 times longer


High
Low
Model Output / Suspicion level
Prevented Activity
Fraud First Identified with model
Fraud First Identified in normal scenario
Duration
Model Close Date

- This would result in an optimal allocation of resources to appropriate claims.

What Exactly is Predictive Modeling?

Techniques of Predictive Modeling

Supervised techniques:

- Statistical Modeling : Build a model for rare events based on labeled data and use it to classify each event

Pros : They produce models that can be easily understood and are easy to implement.

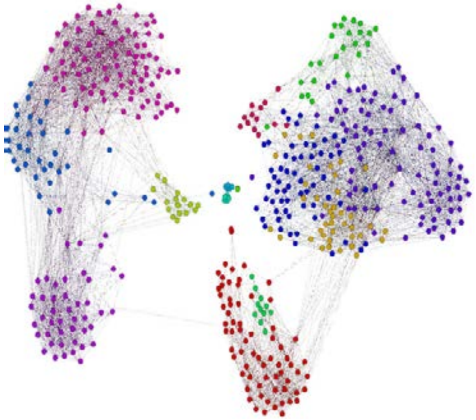
Cons : Statistical Modeling on rare events can lead to inaccurate results

Unsupervised techniques:

- Involves analysis of each event to determine how similar (or dissimilar) it is to the majority
- Stochastic modeling
- Clustering
- Other association rules

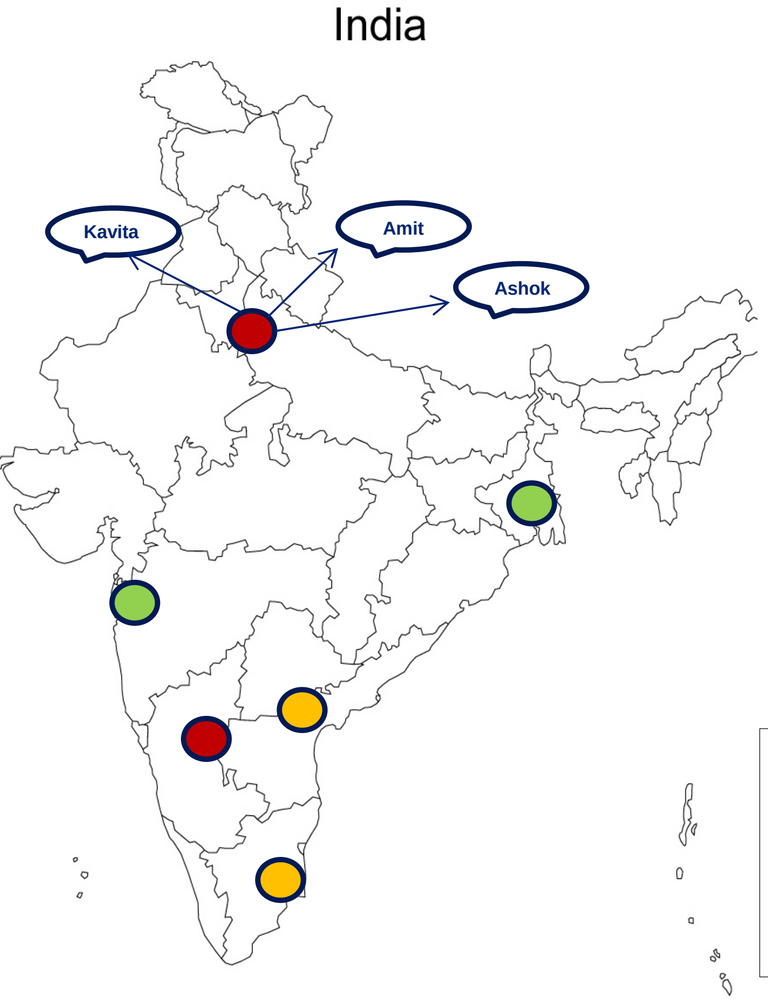
Pros : Can be applied to rare events

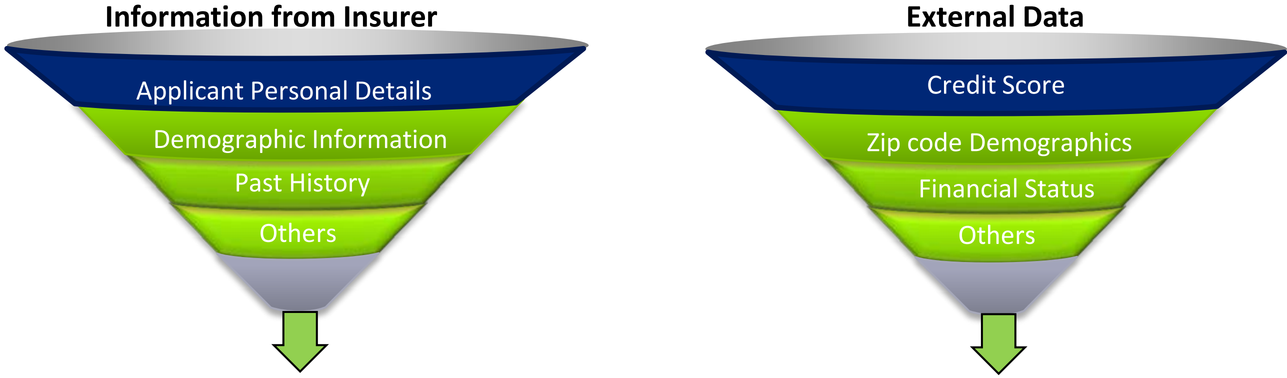
Cons : Tough to identify the exhaustive list of all fraud cases



Deloitte.

Supervised Techniques : Statistical Modeling





Information from Insurer

Applicant Personal Details

Demographic Information

Past History

Others

External Data

Credit Score

Zip code Demographics

Financial Status

Others

Kavita

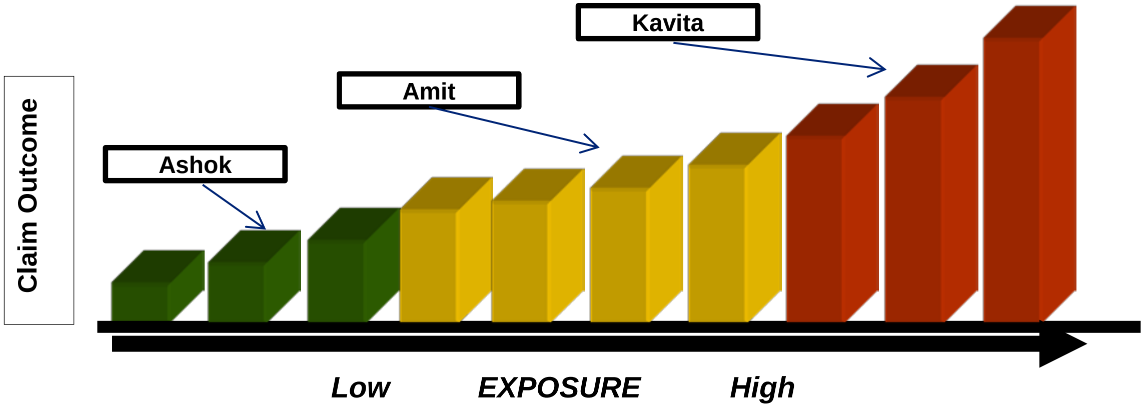
Age : 30
Monthly Income
18,000 INR
Marital Status :
Married
Children : 2
Spouse : Non - working
High crime rate in location
Credit score: Low

Amit

Age : 32
Monthly Income
28,000 INR
Marital Status :
Married
Children : 2
Spouse : Non - working
Average crime rate in location
Credit score: Medium

Ashok

Age : 26
Monthly Income
32,000 INR
Marital Status :
Married
Children : None
Spouse : working
Low crime rate in location
Credit score: High



Deloitte.

Unsupervised Techniques : Outlier Detection

Definition

- Outlier is defined as a data point which is very different from the rest of the data, based on some pre-determined measure
- Outliers can be identified using either
  1. Statistics based approaches (Example: Stochastic modeling)
  2. Distance based approaches (Example: Clustering)

Steps Involved

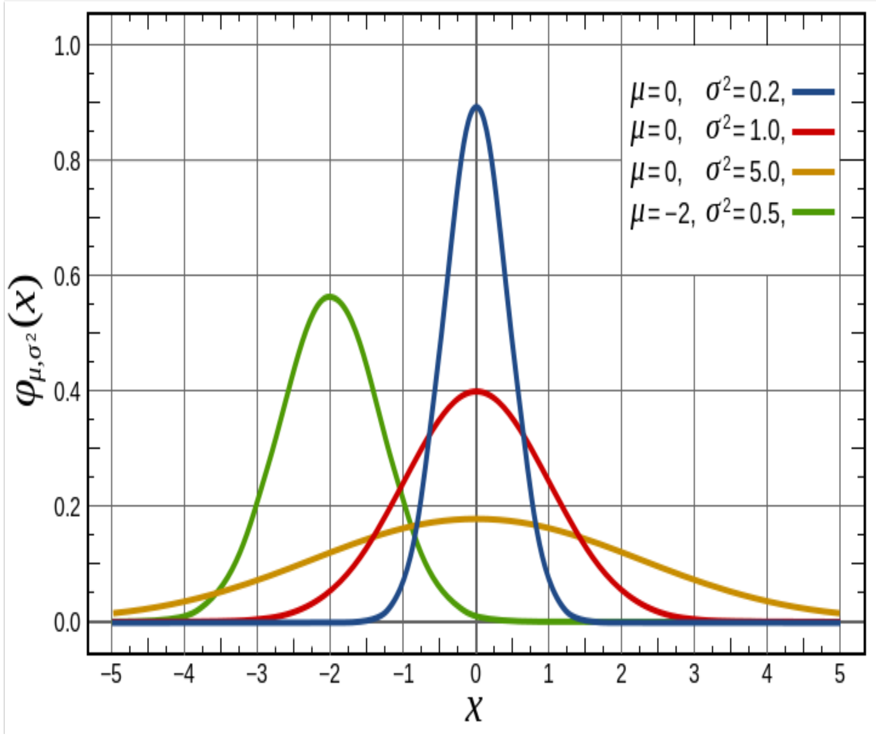
- Identification of Normal Behavior
- Define Similarity Function
- Outlier Detection Algorithm

Pros & Cons

- Pros : Fraud events are usually different from non-fraud events and this type of detection will certainly help in capturing the fraud events
- Cons : This kind of approach could give rise to high false alarm rate – previously unseen yet legitimate data records may now be identified as fraud

Outlier Detection – Statistics based Approaches

- The main underlying assumption is that the number of normal elements is more than the number of outliers
- Data points are modeled using a stochastic distribution (based on prior information) and can be represented with Histogram density or finite mixture model depending on the type of variable (categorical or stochastic)
- We detect the outliers on the basis of their relationship with above model. For this, a likelihood function is used to calculate the probability of each point being an outlier



- Suppose there are 4 groups of data points each following a different univariate distribution as shown.
- Say, a new data point is to be classified into one of the 4 groups.
- One can calculate the p-value associated with each of the 4 distributions and the point goes into that distribution that gives the highest p-value.

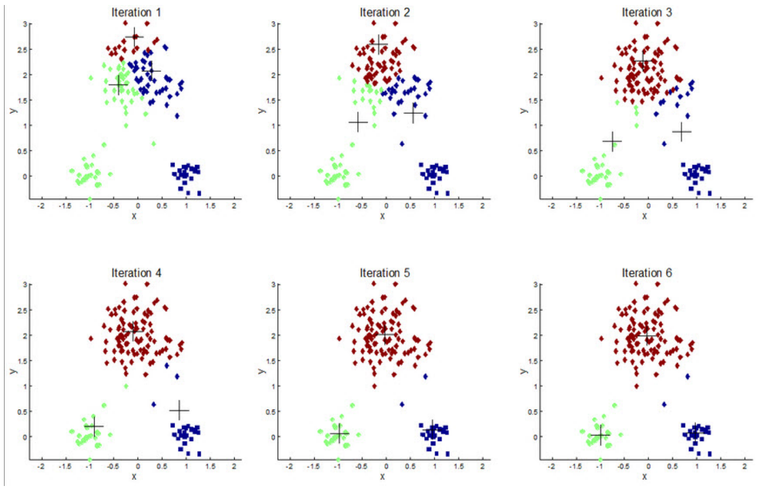
Distance based Outlier Detection Schemes

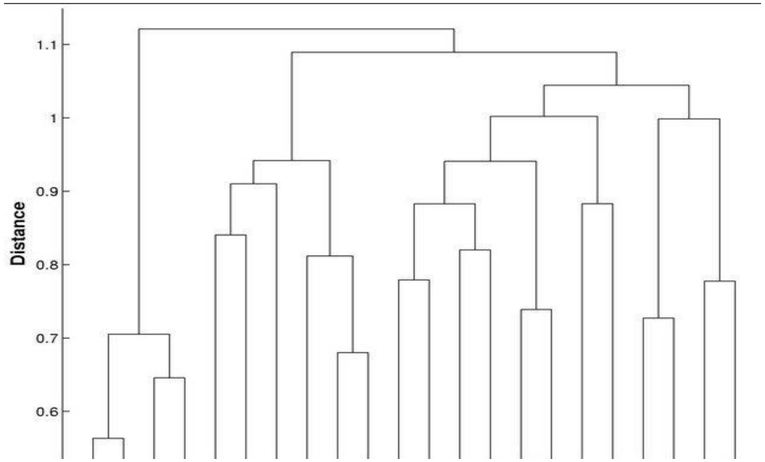
Clustering:

- Clustering is a task of grouping a set of objects in such a way that objects in the same cluster are more similar to each other than those in other clusters
- For each data point, the distance to the nearest neighbor is computed and outliers located in the most sparse neighborhoods are identified based on the distance measure (Examples: Euclidean distance, Mahalanobis distance and Manhattan distance
- Clustering techniques are distinguished based on the type of algorithm used to identify nearest neighbors. Different types of clustering are:

K – means Clustering – Start with K different means and classify every data point into each of these K clusters using nearest neighbor technique.

Hierarchical Clustering – Start with all the data points in different clustering and keep grouping the data points together to obtain the required number of clusters.





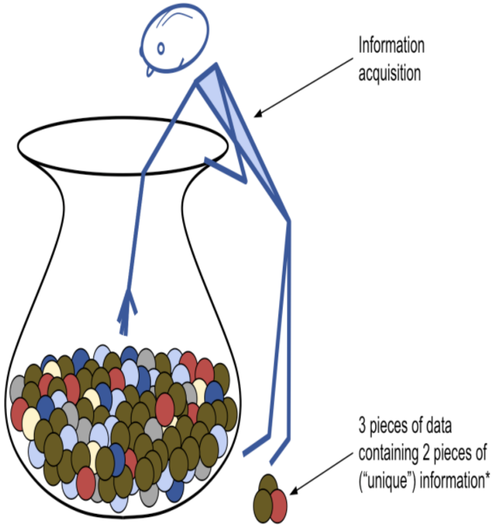
Challenges

Identification Problem:

- Many fraudulent claims often remain undetected and hence be termed as legitimate. The characteristics of such claims are similar to the other detected fraudulent claims and thus make it difficult to identify fraud
- Outlier detection techniques help in solving this kind of problem to some extent

Rare events:

- Rare events are events that occur very infrequently, i.e., their frequency ranges from 0.1% to less than 10%. However, when they do occur, their consequences can be quite dramatic and quite often in negative sense
- Millions of regular transactions are stored while only a few of them are actually fraud
- Standard approaches for feature selection and construction do not work well for rare class analysis
- OverSampling is one common technique to deal with rare events data where a sample is usually drawn from the entire population in such a way that the sample is still a representation of the population while at the same time increasing the proportion of fraud cases
- There are different OverSampling techniques and as such there is no one single best approach


Information acquisition
3 pieces of data containing 2 pieces of ("unique") information*

* Data interpreted as redundant representation of information

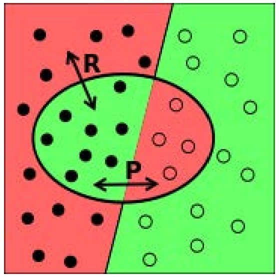
Conclusion

- Supervised techniques will help understand how risky a particular claim is to the business while unsupervised techniques help in identifying the outliers
- But it is not clear as to how a case will be flagged as fraud or how to decide on a particular case as outlier

Classification using F - measure

Confusion matrix:

Predicted Cases (Expectation)	Actual Cases (Observations)	
	Fraud - 1	**Non - Fraud - 0**
Fraud - 1	**TP** (True Positive) Correct Result	**FP** (False Positive) Unexpected Result
Non - Fraud - 0	**FN** (False Negative) Missing Result	**TN** (True Negative) Correct Absence of Results



Precision (P) = TP/(TP + FP)
Recall (R)   = TP/(TP + FN)
F – measure = 2*P*R/(P+R)

- **Recall is the ratio between the number of correctly detected fraud cases and the total number of fraud cases**
- **Precision is the ratio between the number of actual fraud cases and the total number of fraud cases detected by the model**
- **F – measure is a trade – off between Precision and Recall. The cut-off value that gives the highest F – score is chosen as the optimal cut-off**

Conclusion – Cont.

Cost Sensitive Classification:

- Misclassification is often associated with some cost. The cost of classifying a fraud case as non-fraud is the value of the fraud while classifying a non-fraud case as fraud will have some investigation charges associated.

  - Let ***C***(i,j) denote the cost of predicting class i as class j
  - ***M***(i,j) denote the number of observations predicted as class j when they should have been in class i
  - Expected misclassification cost is the Hadamard product of ***M*** and ***C*** divided by the number of observations (N)

$$\frac{1}{N}\sum_{i,j=1}^{N} M(i,j).C(i,j)$$

Predicted Cases (Expectation)	Actual Cases (Observations)	
	Fraud - 1	**Non - Fraud - 0**
Fraud - 1	**C(1,1) = 0**	**C(1,0) = 1 INR**
Non - Fraud - 0	**C(0,1) = 100 INR**	**C(1,1) = 0**

There is no one correct or complete approach to detect fraud but the best way is to use a combination of Supervised and Unsupervised techniques to maximize the prediction power and thus prevent Fraud.

QUESTIONS???