Deloitte.

Fraud in Insurance

Applications of Predictive Modeling

Debashish Banerjee

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)

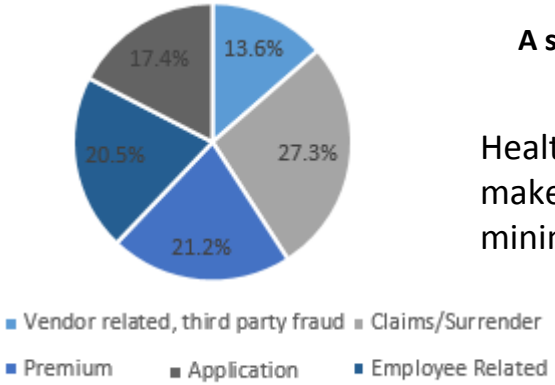
Insurance fraud occurs when any act is committed with the intent to fraudulently obtain some benefit or advantage to which they are not otherwise entitled or someone knowingly denies some benefit that is due and to which someone is entitled.

TRIVIA

According to a recent survey by insurance institute of India, it is estimated that the number of false claims in the Indian industry is approximately 15 per cent of total claims

The same report suggests that the healthcare industry in India is losing approximately Rs.600-Rs 800 crores incurred on fraudulent claims annually.

Fraud risk exposure faced by insurance companies



A survey shows that maximum fraud risk exposure is in the area of claims

Health insurance is a bleeding sector with very high claims ratio. Hence, in order to make health insurance a viable sector, it is essential to concentrate on elimination or minimization of fake claims.

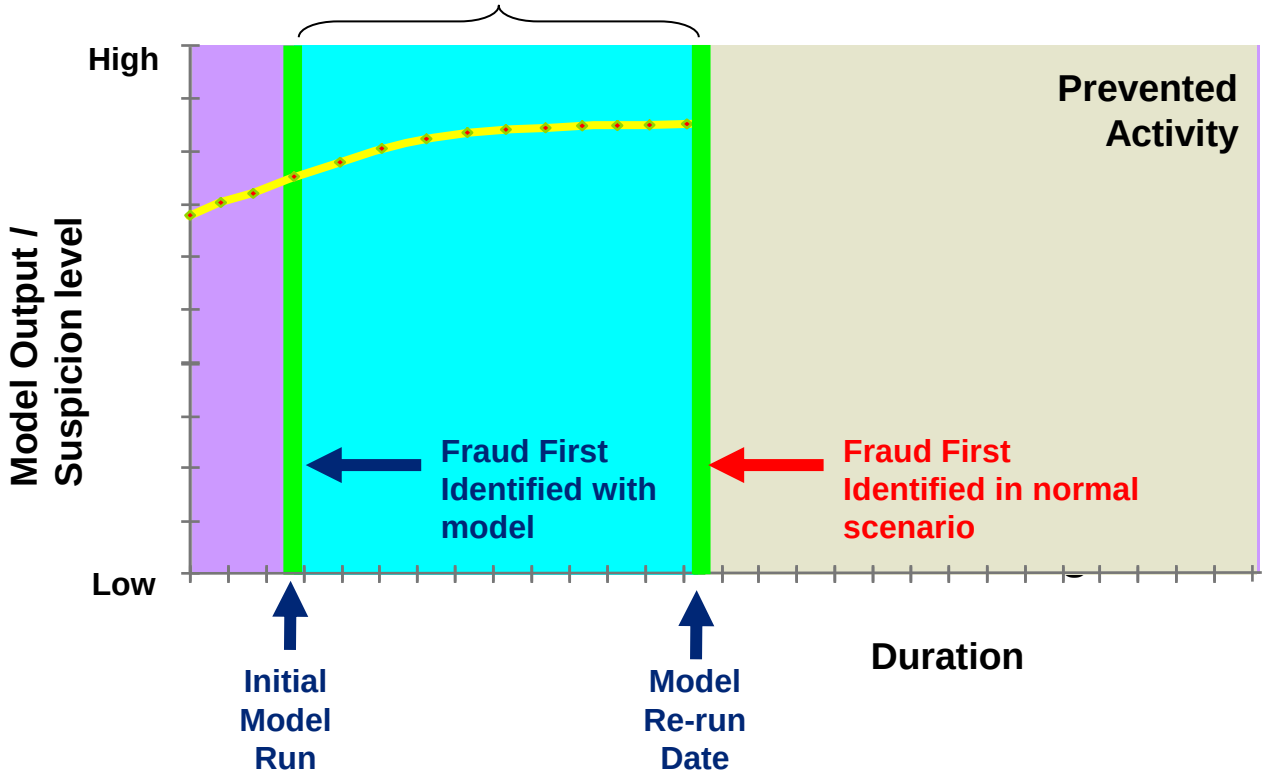
Deloitte.

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.

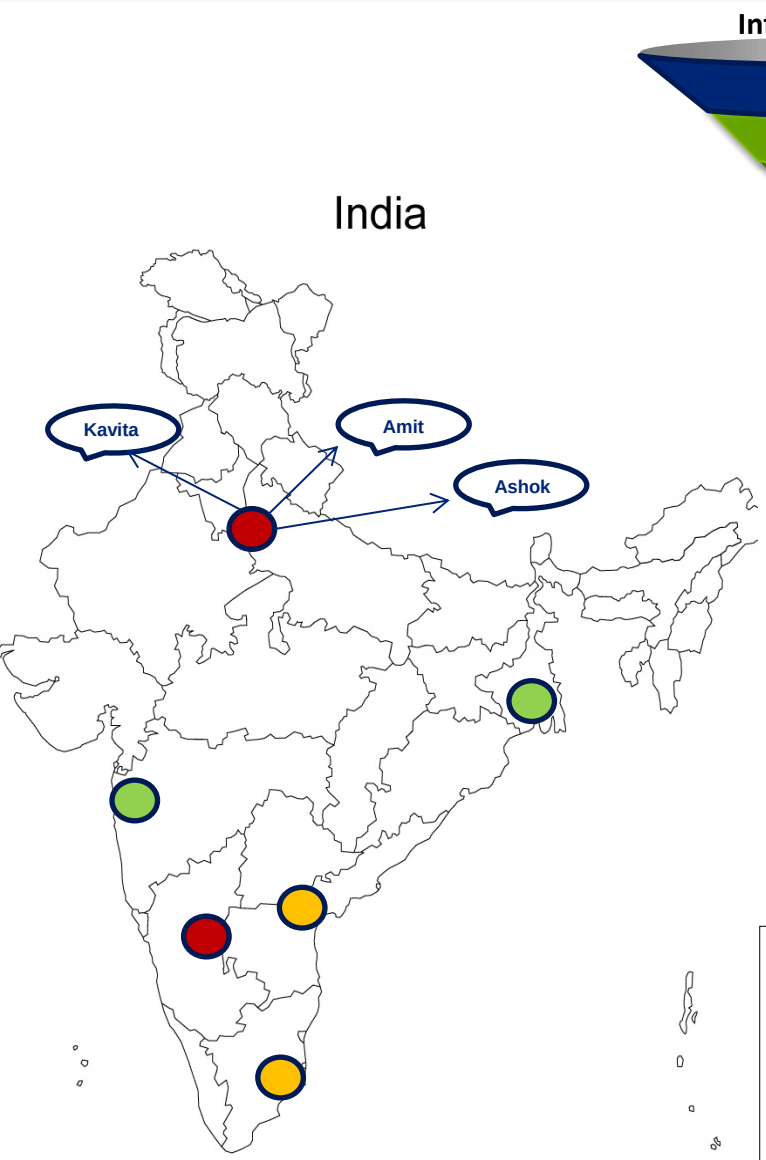
Can Predictive Modeling be the Answer?

- 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

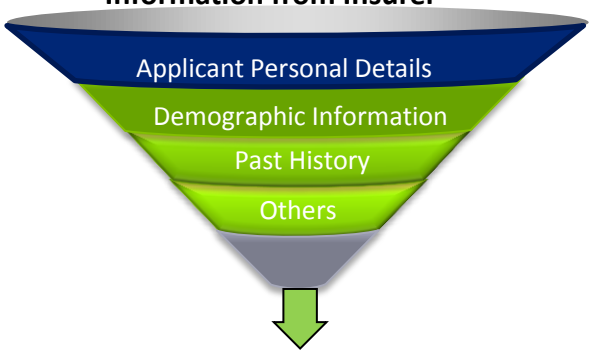


- This would result in an optimal allocation of resources to appropriate claims.

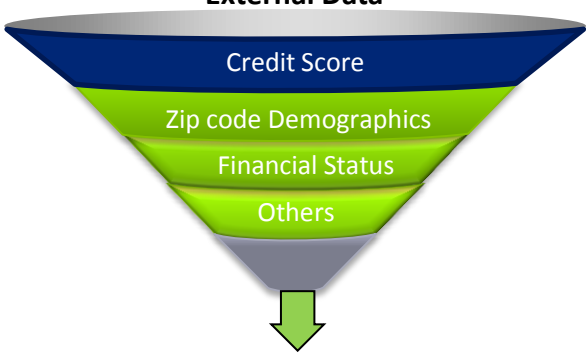
Fraud Analytics Implementation



Information from Insurer



External Data



Ashok

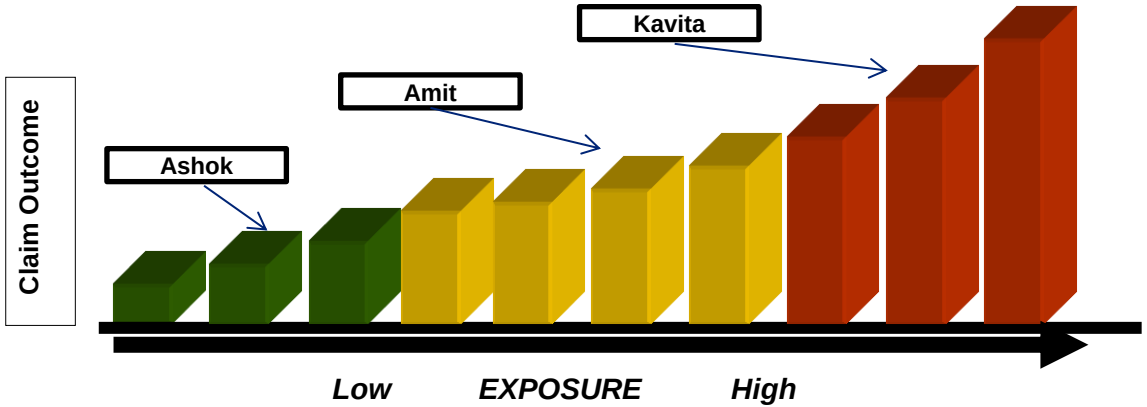
Age : 26
Monthly Income 32,000 INR
Marital Status : Married
Children : None
Spouse : working
Low crime rate in location
Credit score: High

Amit

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

Kavita

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



Challenges in Detecting Fraud

Identification issues

- Use of Unsupervised techniques like Association rules and outlier detection technique.

Rare events – Model could be Biased

- Supervised techniques such as Over Sampling? Bootstrapping ?

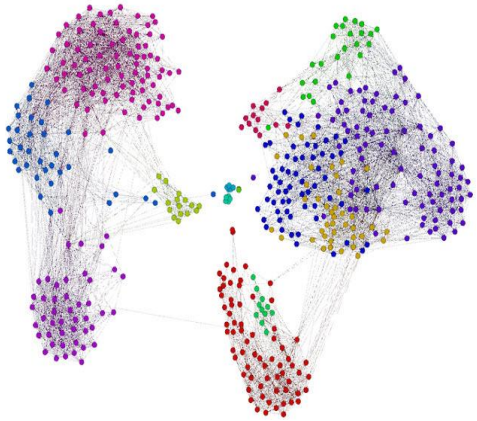
Techniques of Predictive Modeling

Unsupervised techniques:

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

Pros : Can be applied to rare events

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



Supervised techniques:

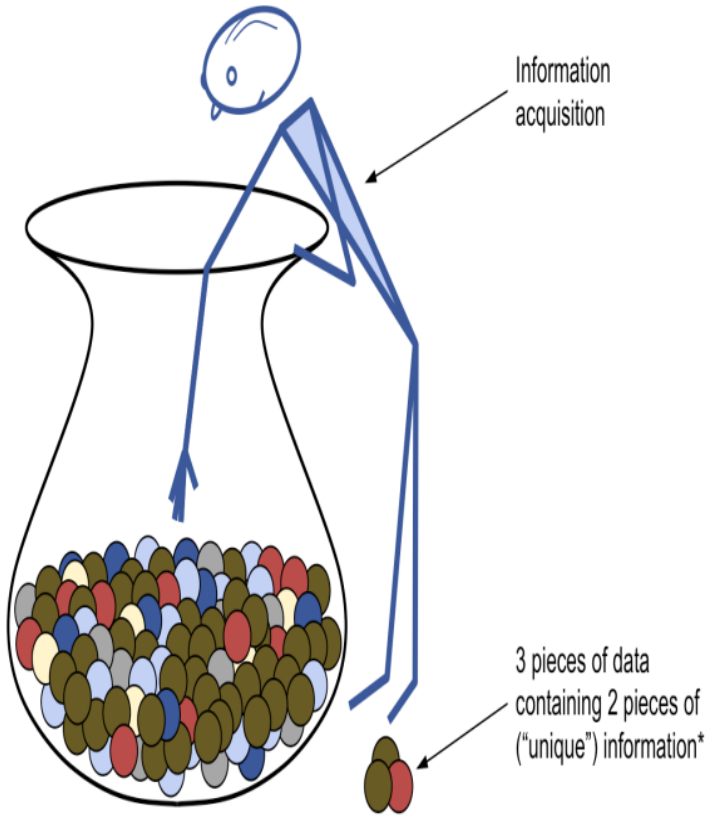
- Regression / Logistic / Probit Modeling
- Statistical Modeling : Build a model for rare events based on Oversampled Data and use it to classify each event

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

➢ 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 e.g. simple random, stratified, bootstrapping etc. but as such there is no one single best approach



* Data interpreted as redundant representation of information

References :

1.JianxinWu James M. Rehg Matthew D. Mullin - Learning a Rare Event Detection Cascade by Direct Feature Selection

2. Aleksandar Lazarević, Jaideep Srivastava, Vipin Kumar - Data Mining for Analysis of Rare Events: A Case Study in Security, Financial and Medical Applications

3. PewResearchCenter U.S. Politics & Policy – Oversampling

4. G. Naga Satish, Ch. V. Raghavendran, Prof. P. Suresh Varma – Intrusion detection and Prevention in Wireless Adhoc Networks

5. J. Miao,*² J. Kirz and D. Sayre - The oversampling phasing method

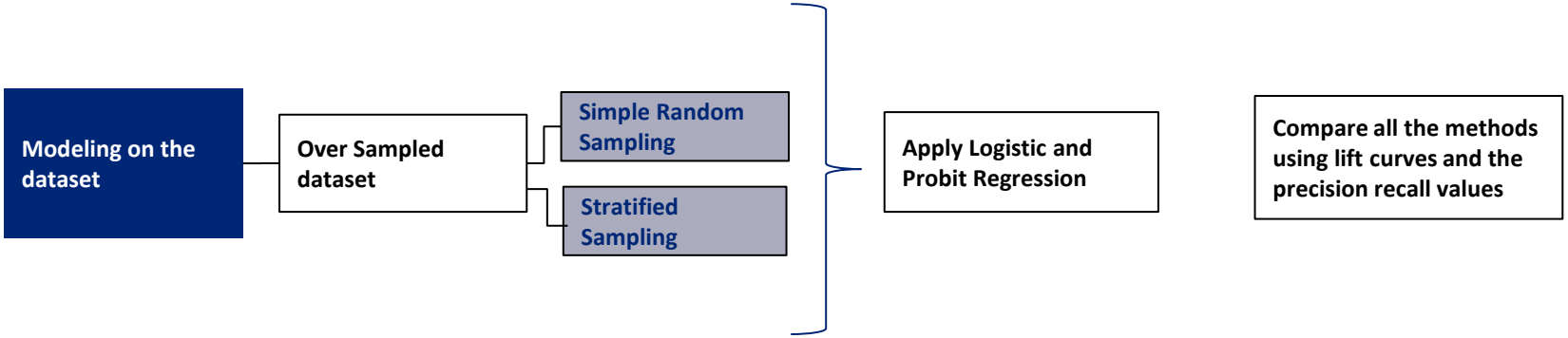
Summary of data:

➢In this example, we talk about fraud in the unemployment insurance sector. The Department of Labor's Unemployment Insurance (UI) programs provide unemployment benefits to eligible workers who become unemployed through no fault of their own, and meet certain other eligibility requirements.

- Data corresponding to unemployment accounts created during the period 2009 – 2010 has been considered.
- The period accounted for about 550,000 unique applicants of whom, approximately 20,000 have been identified as fraud using some cross-matching against employer filings. In other words, data has about 3% cases flagged as fraud and 97% flagged as non-fraud.
- We assume that all the cases flagged as fraud in the model are fraud and all cases flagged as non-fraud are not fraud.
- The data received includes information about
  - ✓ Applicant – age, gender, race, etc.,
  - ✓ Work history – industry, Occupation etc.,
  - ✓ Account – Date of application, benefits details, etc.,
  - ✓ Applicant history - # past accounts, # past frauds, total benefits paid etc.,
  - ✓ Indicator for Overpayment
- The same set of independent variables have been used for all the models developed to start with.

- The following method has been used to construct models
  - ✓ Divide the whole dataset into 2 parts:
    - Modeling dataset – Consists of the first 18 months of data
    - Validation dataset – Consists of the last 6 months of data
  - ✓ Models are built on the modeling dataset or variants of the modeling dataset (Obtained from Oversampling) and tested on the validation dataset

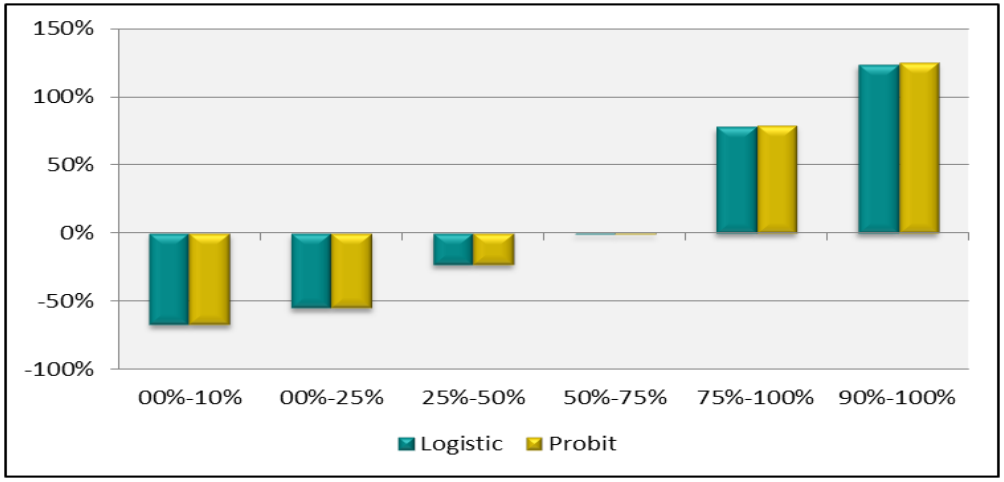
- For comparing the model performances, we would be using two metrics:
  - Lift Curves
  - Precision Recall values



Lift Curves and the correlation matrix show that segmentation is not affected by the sampling scheme or the regression technique and that inferences can be made using the entire population directly.

Oversampling – Simple Random Sample

- Oversampling has been performed as follows:
  - ✓ Select all the fraud cases in the modeling dataset into the sample
  - ✓ Randomly select thrice as many non-fraud cases from the modeling dataset as there are fraud cases in the sample and obtain the resampled dataset
- Two models have been built on the resampled dataset, one using logistic regression and the other using Probit regression
- A comparison of these two models is as follows:



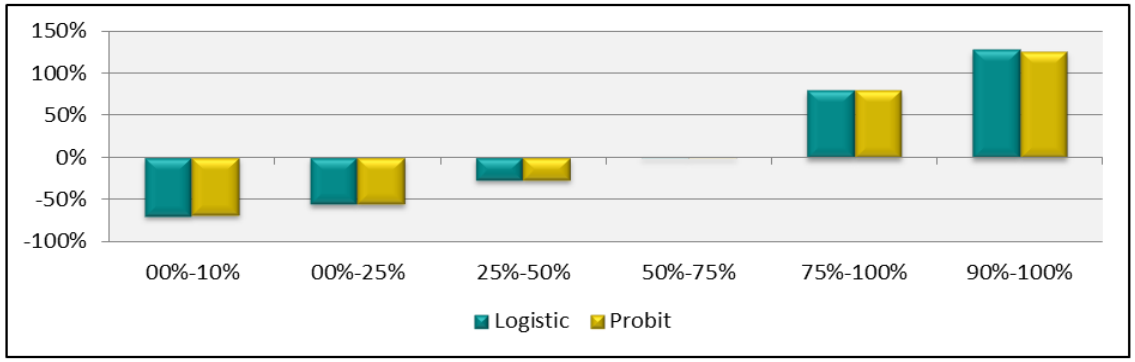
	Num_fraud	
Decile	Logistic	Probit
00%-10%	120	121
10%-20%	176	180
20%-30%	230	224
30%-40%	291	295
40%-50%	284	282
50%-60%	309	307
60%-70%	378	386
70%-80%	443	432
80%-90%	575	575
90%-100%	809	813

- Logistic denotes the fraud ratio relativity values obtained in Logistic regression
- Probit denotes the fraud ratio relativity values obtained in Probit regression
- Fraud ratio relativity is obtained by dividing the difference between the average fraud ratio of the decile and the overall average fraud ratio by the overall average fraud ratio.

Oversampling – Stratified Sampling - Continued

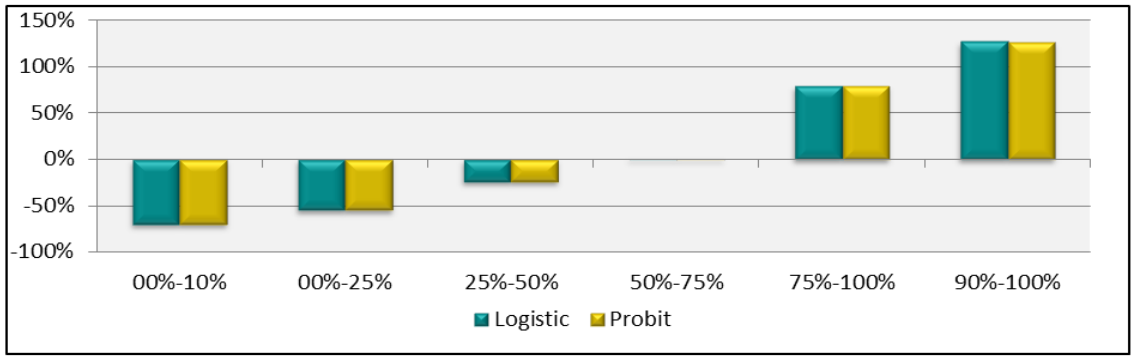
The lift curves constructed for variations 1 and 2 are as follows:

Variation 1: Fraud ratio = 3



	Num_fraud	
Decile	Logistic	Probit
00%-10%	112	115
10%-20%	195	193
20%-30%	211	222
30%-40%	287	279
40%-50%	271	271
50%-60%	315	310
60%-70%	379	377
70%-80%	458	459
80%-90%	562	573
90%-100%	825	816

Variation 2: Fraud ratio = 1



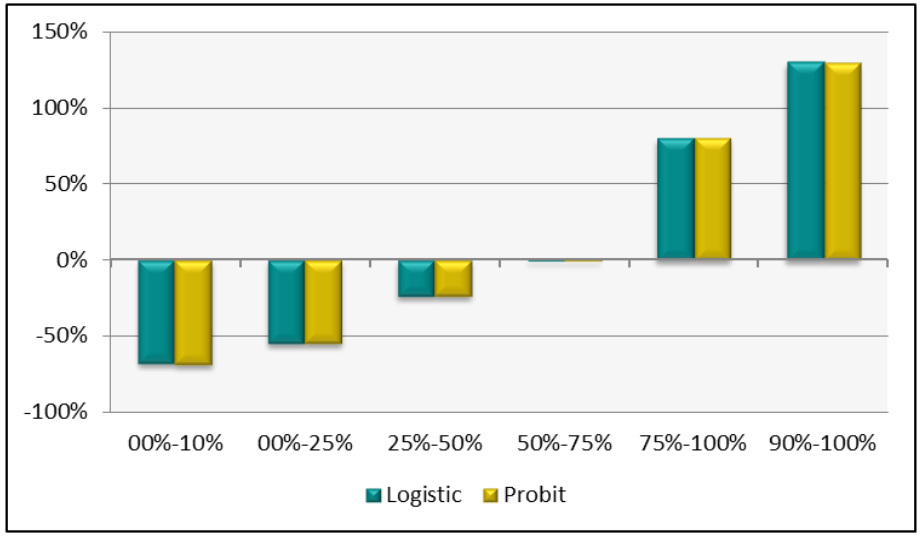
	Num_fraud	
Decile	Logistic	Probit
00%-10%	109	109
10%-20%	199	198
20%-30%	218	217
30%-40%	273	273
40%-50%	297	298
50%-60%	307	306
60%-70%	358	362
70%-80%	481	479
80%-90%	554	555
90%-100%	819	818

- Logistic denotes the fraud ratio relativity values obtained in Logistic regression
- Probit denotes the fraud ratio relativity values obtained in Probit regression
- Fraud ratio relativity is obtained by dividing the difference between the average fraud ratio of the decile and the overall average fraud ratio by the overall average fraud ratio.

Modeling on the Entire Dataset

- Two models, one using Logistic regression and the other using Probit regression have been built on the modeling dataset and have been tested on the Validation dataset.

- A comparison of the lift curves for the two models is as follows:



	Num_fraud	
Decile	Logistic	Probit
00%-10%	117	112
10%-20%	192	195
20%-30%	214	222
30%-40%	282	278
40%-50%	283	286
50%-60%	304	302
60%-70%	373	358
70%-80%	459	468
80%-90%	561	565
90%-100%	830	829

- Logistic denotes the fraud ratio relativity values obtained in Logistic regression
- Probit denotes the fraud ratio relativity values obtained in Probit regression
- Fraud ratio relativity is obtained by dividing the difference between the average fraud ratio of the decile and the overall average fraud ratio by the overall average fraud ratio.

Conclusions

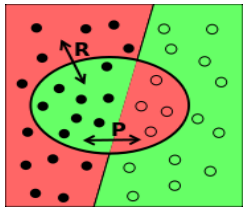
- All the sampling schemes suggest that logistic and Probit produce similar segmentation results.
- Observations have been ranked based on the probability of being fraud in each of the 4 methods. A look at the rank correlation coefficients between the estimated probabilities for all the 4 methods is as follows:

CORR MATRIX	Model 1	Model 2	Model 3	Model 4
Model 1	1	1	1	0.99
Model 2	1	1	0.99	0.99
Model 3	1		1	0.99
Model 4	0.99	0.97	0.97	1

- Model 1 - Model constructed by performing Logistic Regression on the entire dataset
- Model 2 - Model constructed by performing Logistic Regression on Simple Random Sample (SRS)
- Model 3- Model constructed by performing Logistic Regression on Stratified Sample (Variation1)
- Model 4- Model constructed by performing Logistic Regression on Stratified Sample (Variation2)

- All the correlations are very high, usually around 99.8%. This suggests that the ranking of observations based on different methods remains pretty much the same irrespective of the method we employ.

Conclusions…Continued



Recall (R) = TP/(TP + FN)
Precision (P) = TP/(TP + FP)
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 correctly detected 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**

➢For every method, the cut-off value beyond which applicants are flagged as fraud is determined and precision recall values have been calculated at these cut-offs. A comparison of precision and recall values obtained in each of the methods is as follows:

Sampling Scheme	Cut off	Precision	Recall	F-Score
Entire Dataset	0.062	7.20%	23.60%	0.11
SRS	0.35	7%	24.00%	0.11
Stratified(Variation 1)	0.36	7.20%	21.50%	0.11
Stratified(Variation 2)	0.37	7%	23.50%	0.11

➢All these show that segmentation is not affected by the sampling scheme or the regression technique and that inferences can be made using the entire population directly.

Questions???