FITE7410

Lecture 2

Lecturers: Dr. Vivien CHAN, Annie CHAN
Tutor: Ms. Yanan GONG

Department of Computer Science
The University of Hong Kong

Today's Agenda

01 **Exploratory Data Analysis (EDA) using R**

02 **Financial Statement Fraud Scheme**

03 **Imbalance Data Handling**

04 **ML Algorithm - Linear and Logistic Regression**

05 **Performance Evaluation**

01
Exploratory Data Analysis (EDA) using R

Dr. Vivien CHAN

Data Visualization with R

Rob Kabacoff (2020)

https://rkabacoff.github.io/datavis/index.html

Example of EDA using R

- Dataset : https://www.kaggle.com/mlg-ulb/creditcardfraud
- The Credit Card Fraud Detection Dataset comprises transactions that European credit card holders made in September 2013. The dataset shows transactions that occurred in two days.
- The dataset has been collected and analyzed during a research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection.

What are the goals of EDA?

How to achieve the goals?

Example – Loading packages and libraries

```r
#Loading packages and libraries

install.packages('corrplot')
library(corrplot)

install.packages('caret')
library(caret)
```

#library for correlations

#library for plotting the samples

```r
library(tidyverse) # metapackage of all tidyverse packages
```

```r
#load csv dataset
data <- read.csv('../input/creditcardfraud/creditcard.csv')
```

Path of the source file

Example - Step1: Distinguish Attributes

```r
#show structure of the dataset
print("Structure of dataset")
str(data)
```

```
[1] "Structure of dataset"
'data.frame':	284807 obs. of  31 variables:
 $ Time  : num  0 0 1 1 2 2 4 7 7 9 ...
 $ V1    : num  -1.36 1.192 -1.358 -0.966 -1.158 ...
 $ V2    : num  -0.0728 0.2662 -1.3402 -0.1852 0.8777 ...
 $ V3    : num  2.536 0.166 1.773 1.793 1.549 ...
 $ V4    : num  1.378 0.448 0.38 -0.863 0.403 ...
 $ V5    : num  -0.3383 0.06 -0.5032 -0.0103 -0.4072 ...
 $ V6    : num  0.4624 -0.0824 1.8005 1.2472 0.0959 ...
 $ V7    : num  0.2396 -0.0788 0.7915 0.2376 0.5929 ...
 $ V8    : num  0.0987 0.0851 0.2477 0.3774 -0.2705 ...
 $ V9    : num  0.364 -0.255 -1.515 -1.387 0.818 ...
 $ V10   : num  0.0908 -0.167 0.2076 -0.055 0.7531 ...
 $ V11   : num  -0.552 1.613 0.625 -0.226 -0.823 ...
 $ V12   : num  -0.6178 1.0652 0.0661 0.1782 0.5382 ...
 $ V13   : num  -0.991 0.489 0.717 0.508 1.346 ...
 $ V14   : num  -0.311 -0.144 -0.166 -0.288 -1.12 ...
 $ V15   : num  1.468 0.636 2.346 -0.631 0.175 ...
 $ V16   : num  -0.47 0.464 -2.89 -1.06 -0.451 ...
 $ V17   : num  0.208 -0.115 1.11 -0.684 -0.237 ...
 $ V18   : num  0.0258 -0.1834 -0.1214 1.9658 -0.0382 ...
 $ V19   : num  0.404 -0.146 -2.262 -1.233 0.803 ...
 $ V20   : num  0.2514 -0.0691 0.525 -0.208 0.4085 ...
 $ V21   : num  -0.01831 -0.22578 0.248 -0.1083 -0.00943 ...
 $ V22   : num  0.27784 -0.63867 0.77168 0.00527 0.79828 ...
 $ V23   : num  -0.11 0.101 0.909 -0.19 -0.137 ...
 $ V24   : num  0.0669 -0.3398 -0.6893 -1.1756 0.1413 ...
 $ V25   : num  0.129 0.167 -0.328 0.647 -0.206 ...
 $ V26   : num  -0.189 0.126 -0.139 -0.222 0.502 ...
 $ V27   : num  0.13356 -0.00898 -0.05535 0.06272 0.21942 ...
 $ V28   : num  -0.0211 0.0147 -0.0598 0.0615 0.2152 ...
 $ Amount: num  149.62 2.69 378.66 123.5 69.99 ...
 $ Class : int  0 0 0 0 0 0 0 0 0 0 ...
```

Example - Step1: Distinguish Attributes

```r
#show summary statistics of the dataset
print("Summary statistics")
summary(data)
```

```
[1] "Summary statistics"
      Time              V1                  V2                  V3
 Min.   :     0   Min.   :-56.40751   Min.   :-72.71573   Min.   :-48.3256
 1st Qu.: 54202   1st Qu.: -0.92037   1st Qu.: -0.59855   1st Qu.: -0.8904
 Median : 84692   Median :  0.01811   Median :  0.06549   Median :  0.1799
 Mean   : 94814   Mean   :  0.00000   Mean   :  0.00000   Mean   :  0.0000
 3rd Qu.:139320   3rd Qu.:  1.31564   3rd Qu.:  0.80372   3rd Qu.:  1.0272
 Max.   :172792   Max.   :  2.45493   Max.   : 22.05773   Max.   :  9.3826
       V4                 V5                   V6                  V7
 Min.   :-5.68317   Min.   :-113.74331   Min.   :-26.1605   Min.   :-43.5572
 1st Qu.:-0.84864   1st Qu.:  -0.69160   1st Qu.: -0.7683   1st Qu.: -0.5541
 Median :-0.01985   Median :  -0.05434   Median : -0.2742   Median :  0.0401
 Mean   : 0.00000   Mean   :   0.00000   Mean   :  0.0000   Mean   :  0.0000
 3rd Qu.: 0.74334   3rd Qu.:   0.61193   3rd Qu.:  0.3986   3rd Qu.:  0.5704
 Max.   :16.87534   Max.   :  34.80167   Max.   : 73.3016   Max.   :120.5895
       V8                  V9                  V10                 V11
 Min.   :-73.21672   Min.   :-13.43407   Min.   :-24.58826   Min.   :-4.79747
 1st Qu.: -0.20863   1st Qu.: -0.64310   1st Qu.: -0.53543   1st Qu.:-0.76249
 Median :  0.02236   Median : -0.05143   Median : -0.09292   Median :-0.03276
 Mean   :  0.00000   Mean   :  0.00000   Mean   :  0.00000   Mean   : 0.00000
 3rd Qu.:  0.32735   3rd Qu.:  0.59714   3rd Qu.:  0.45392   3rd Qu.: 0.73959
 Max.   : 20.00721   Max.   : 15.59500   Max.   : 23.74514   Max.   :12.01891
      V12                V13                V14                V15
 Min.   :-18.6837   Min.   :-5.79188   Min.   :-19.2143   Min.   :-4.49894
 1st Qu.: -0.4056   1st Qu.:-0.64854   1st Qu.: -0.4256   1st Qu.:-0.58288
 Median :  0.1400   Median :-0.01357   Median :  0.0506   Median : 0.04807
 Mean   :  0.0000   Mean   : 0.00000   Mean   :  0.0000   Mean   : 0.00000
 3rd Qu.:  0.6182   3rd Qu.: 0.66251   3rd Qu.:  0.4931   3rd Qu.: 0.64882
 Max.   :  7.8484   Max.   : 7.12688   Max.   : 10.5268   Max.   : 8.87774
```

```
      V16                 V17                 V18
 Min.   :-14.12985   Min.   :-25.16280   Min.   :-9.498746
 1st Qu.: -0.46804   1st Qu.: -0.48375   1st Qu.:-0.498850
 Median :  0.06641   Median : -0.06568   Median :-0.003636
 Mean   :  0.00000   Mean   :  0.00000   Mean   : 0.000000
 3rd Qu.:  0.52330   3rd Qu.:  0.39968   3rd Qu.: 0.500807
 Max.   : 17.31511   Max.   :  9.25353   Max.   : 5.041069
      V19                 V20                 V21
 Min.   :-7.213527   Min.   :-54.49772   Min.   :-34.83038
 1st Qu.:-0.456299   1st Qu.: -0.21172   1st Qu.: -0.22839
 Median : 0.003735   Median : -0.06248   Median : -0.02945
 Mean   : 0.000000   Mean   :  0.00000   Mean   :  0.00000
 3rd Qu.: 0.458949   3rd Qu.:  0.13304   3rd Qu.:  0.18638
 Max.   : 5.591971   Max.   : 39.42090   Max.   : 27.20284
       V22                  V23                 V24
 Min.   :-10.933144   Min.   :-44.80774   Min.   :-2.83663
 1st Qu.: -0.542350   1st Qu.: -0.16185   1st Qu.:-0.35459
 Median :  0.006782   Median : -0.01119   Median : 0.04098
 Mean   :  0.000000   Mean   :  0.00000   Mean   : 0.00000
 3rd Qu.:  0.528554   3rd Qu.:  0.14764   3rd Qu.: 0.43953
 Max.   : 10.503090   Max.   : 22.52841   Max.   : 4.58455
      V25                 V26                V27
 Min.   :-10.29540   Min.   :-2.60455   Min.   :-22.565679
 1st Qu.: -0.31715   1st Qu.:-0.32698   1st Qu.: -0.070840
 Median :  0.01659   Median :-0.05214   Median :  0.001342
 Mean   :  0.00000   Mean   : 0.00000   Mean   :  0.000000
 3rd Qu.:  0.35072   3rd Qu.: 0.24095   3rd Qu.:  0.091045
 Max.   :  7.51959   Max.   : 3.51735   Max.   : 31.612198
      V28                Amount             Class
 Min.   :-15.43008   Min.   :    0.00   Min.   :0.000000
 1st Qu.: -0.05296   1st Qu.:    5.60   1st Qu.:0.000000
 Median :  0.01124   Median :   22.00   Median :0.000000
 Mean   :  0.00000   Mean   :   88.35   Mean   :0.001728
 3rd Qu.:  0.07828   3rd Qu.:   77.17   3rd Qu.:0.000000
 Max.   : 33.84781   Max.   :25691.16   Max.   :1.000000
```

Example - Step1: Distinguish Attributes

```
#show label class structure
print("Class labels")
table(data$Class)
print("% of 2 classes")
table(data$Class)/length(data$Class)
```

```
[1] "Class labels"

     0      1
284315    492
[1] "% of 2 classes"

          0           1
0.998272514 0.001727486
```

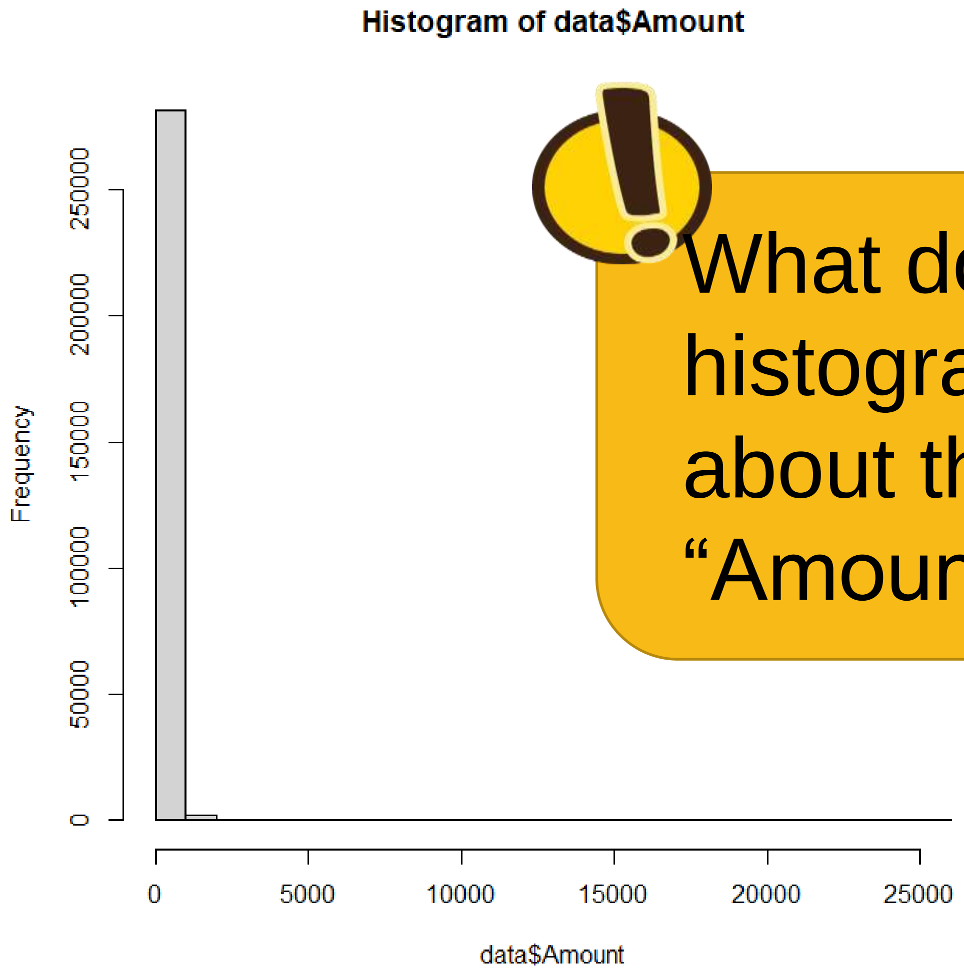
What kind of initial information that you can get from this preliminary exploration?

Example – Step 2: Univariate Analysis

```
help(hist)
```

Use “help” to check the R documentation

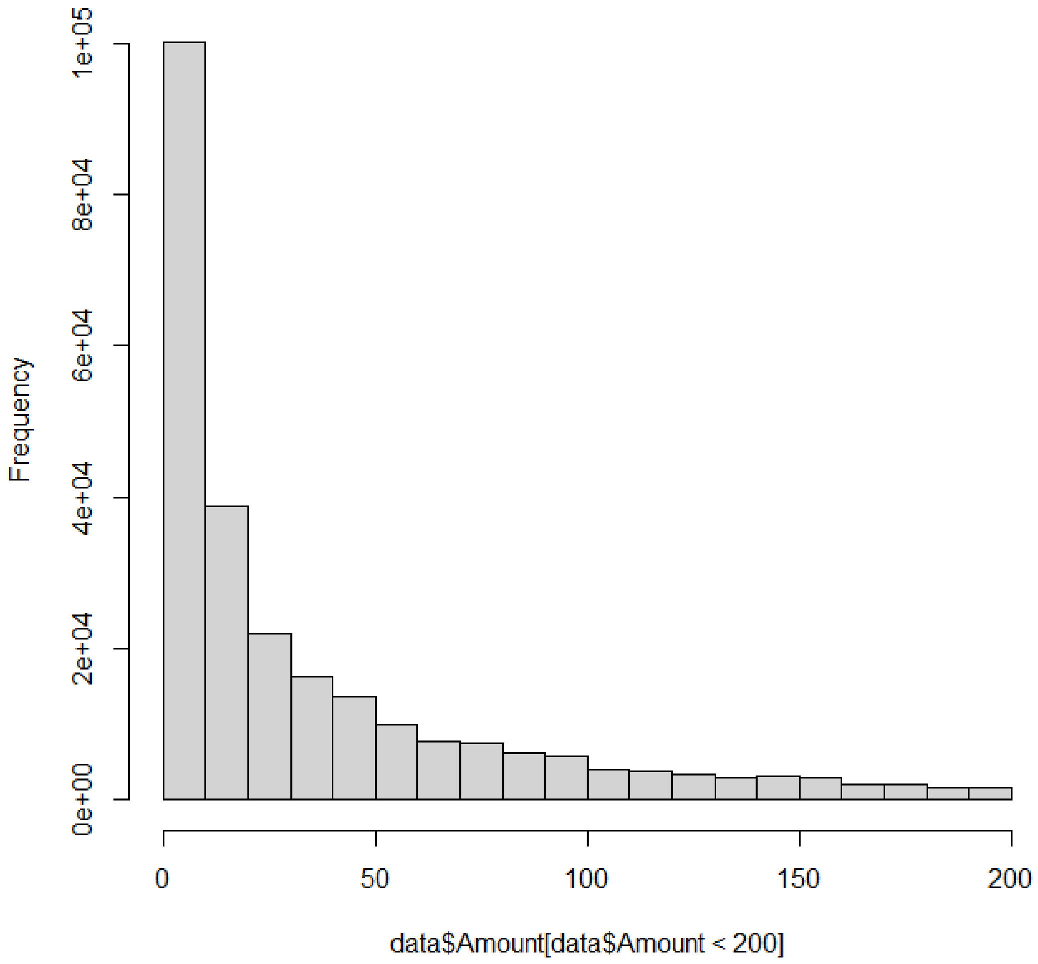
```
#Histogram
hist(data$Amount)
```



What does this histogram tell you about the variable “Amount”?

Example – Step 2: Univariate Analysis

```r
#Plot Histogram for Amount smaller than $200
hist(data$Amount[data$Amount < 200])
```



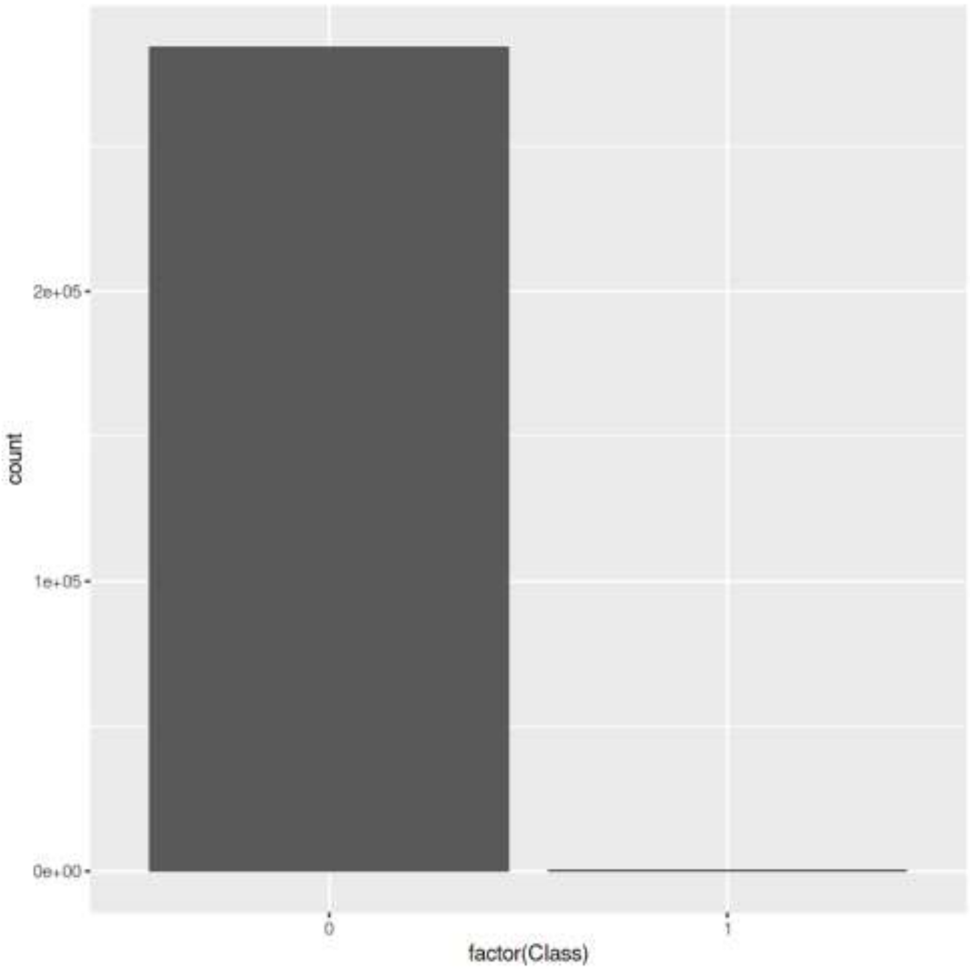
Now, with the “Amount” limited to under $200, what does this histogram tell you about the variable “Amount”?

Example – Step 2: Univariate Analysis

** Examples of bar charts

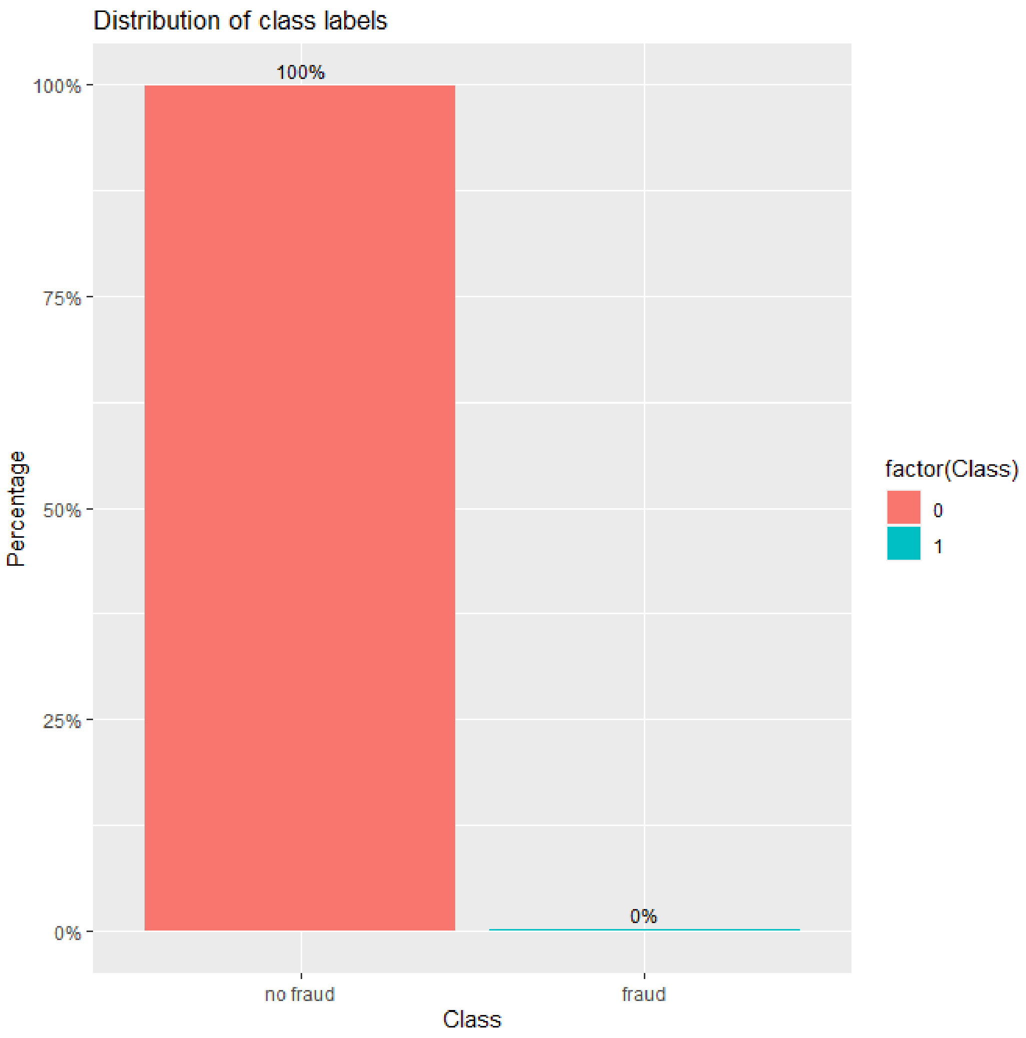
```
#Bar chars plotting using ggplot and geom_bar()
ggplot(data, aes(x = factor(Class))) +
      geom_bar()
```

ggplot() is used to construct the initial plot object, and is almost always followed by + to add component to the plot.



Example – Step 2: Univariate Analysis

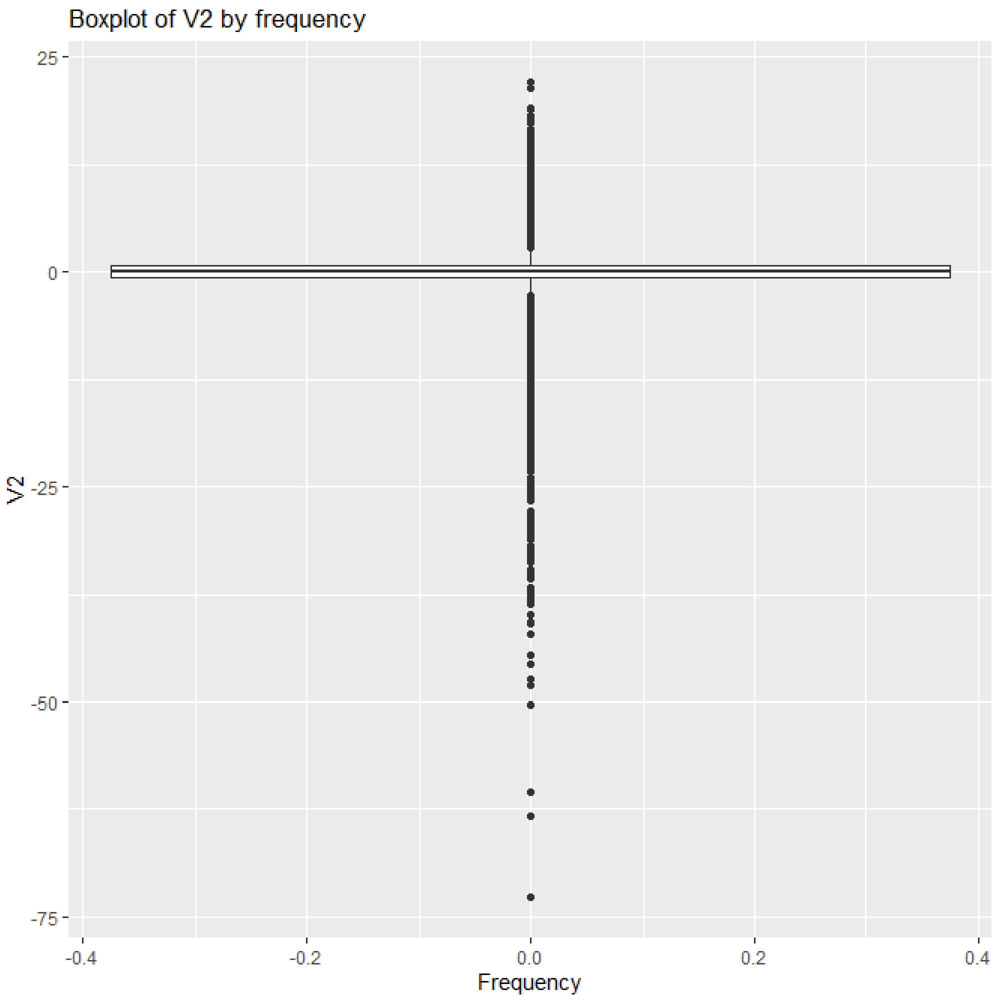
```
#Bar charts
ggplot(data, aes(x = factor(Class),
                 y = prop.table(stat(count)),fill = factor(Class),
                 label = scales::percent(prop.table(stat(count))))) +
    geom_bar(position="dodge") +
    geom_text(stat = 'count',
              position = position_dodge(.9),
              vjust = -0.5,
              size = 3) +
    scale_x_discrete(labels = c("no fraud", "fraud")) +
    scale_y_continuous(labels = scales::percent) +
    labs(x = 'Class', y = 'Percentage') +
    ggtitle("Distribution of class labels")
```



Example – Step 2: Univariate Analysis

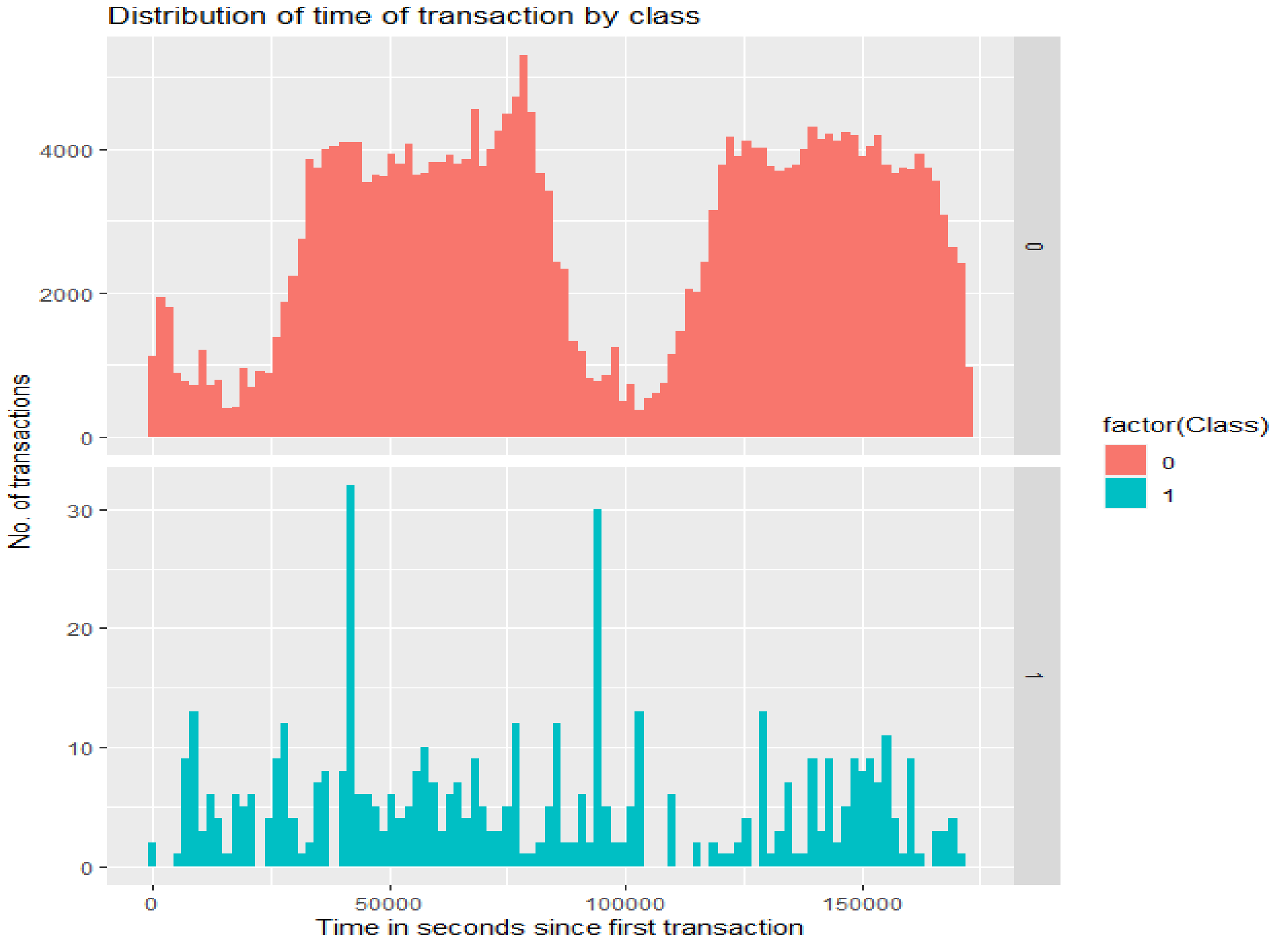
** Examples of boxplot

```
#Boxplot of variable V2
ggplot(data, aes(y=V2)) +
    geom_boxplot() +
    labs(x = 'Frequency', y = 'V2') +
    ggtitle('Boxplot of V2 by frequency')
```



Example – Step 3: Bi-/Multi-variate Analysis

```r
#Histogram plot of variable Time and Class label
ggplot(data, aes(x = Time, fill = factor(Class))) +
    geom_histogram(bins = 100) +
    labs(x = 'Time in seconds since first transaction', y = 'No. of transactions') +
    ggtitle('Distribution of time of transaction by class') +
    facet_grid(Class ~ ., scales = 'free_y')
```



What insights can you get from this figure?

Example – Step 3: Bi-/Multi-variate Analysis

** Examples of plotting correlations

Step 1

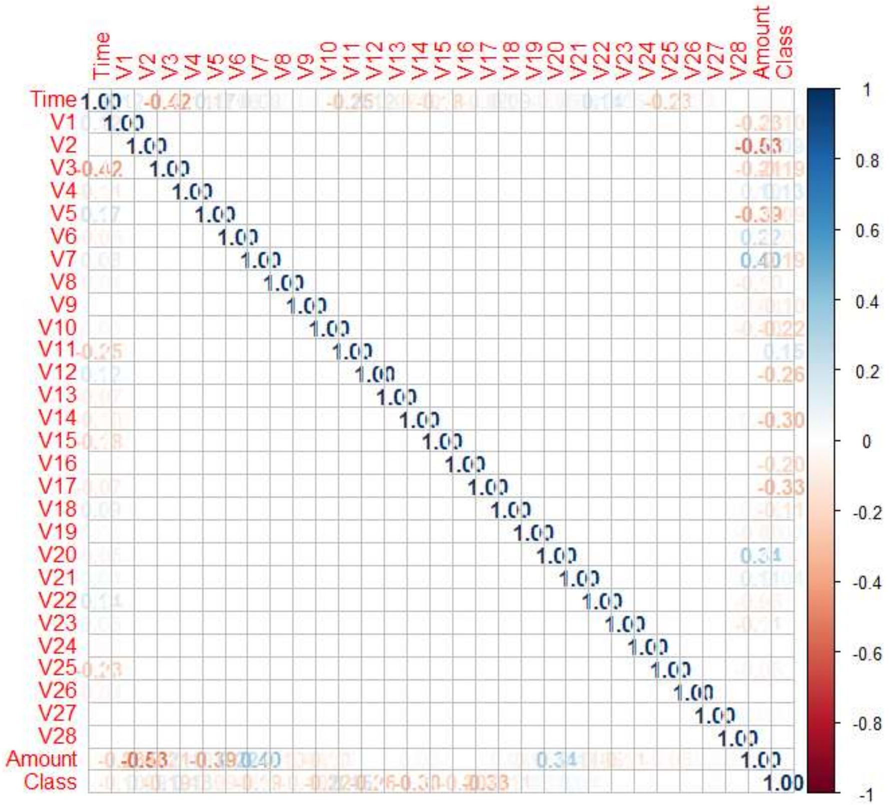
```
#compute the correlations among the variables
corr_mat <- cor(data)
```

Step 2

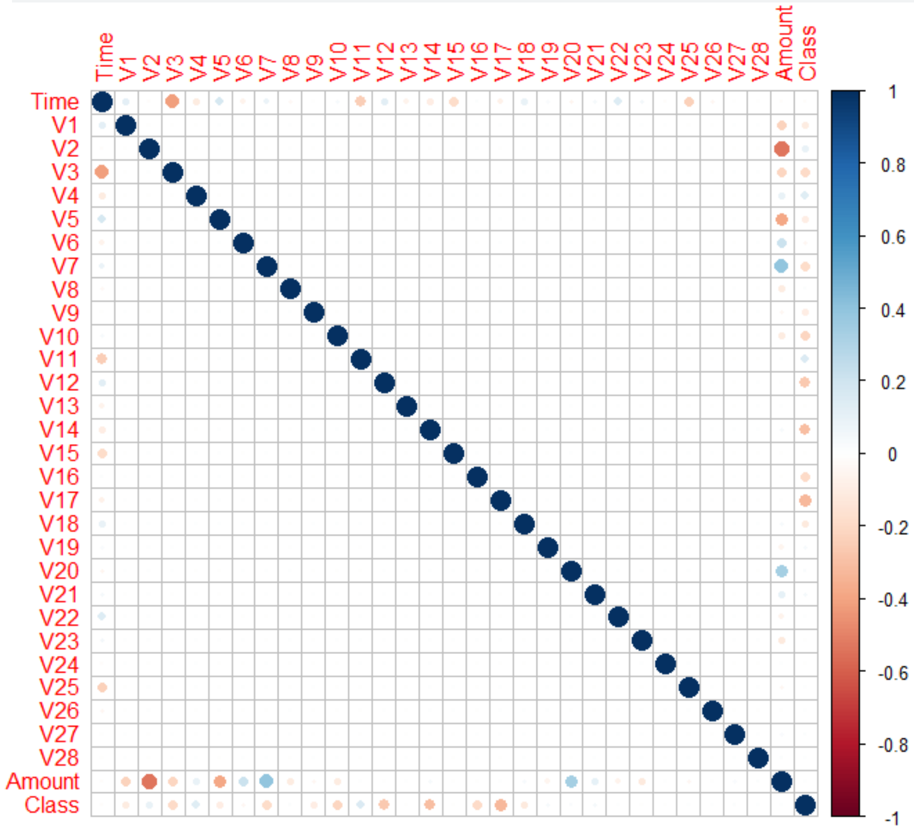
```
#Plot correlation heat map
corrplot(corr_mat, method="number")
corrplot(corr_mat, method="circle")
```

Example – Step 3: Bi-/Multi-variate Analysis

```
corrplot(corr_mat, method="number")
```



```
corrplot(corr_mat, method="circle")
```



What insights can you get from the correlations among the attributes?

02
Financial Statement Fraud Scheme

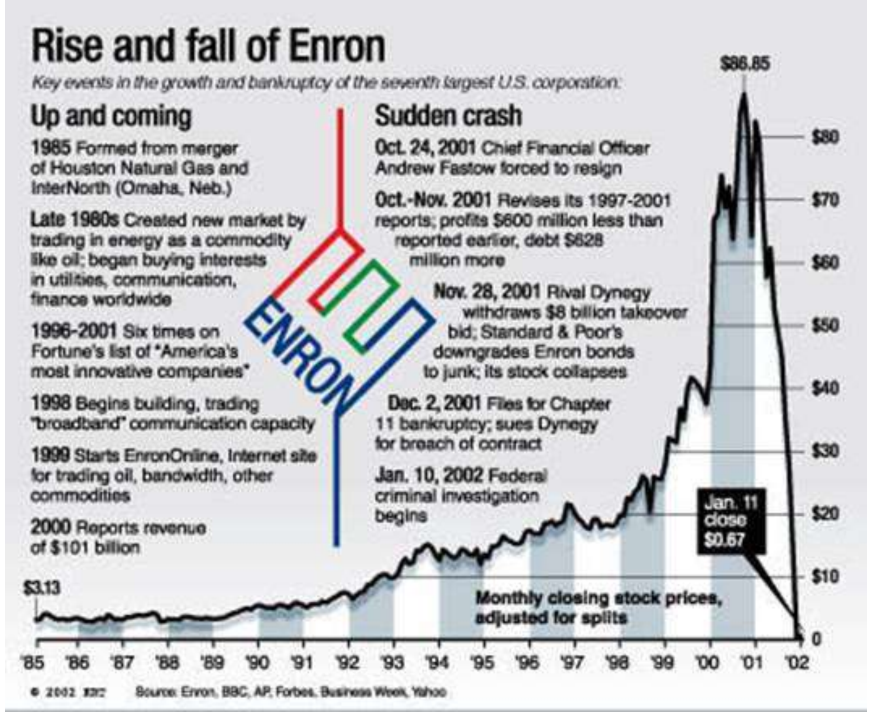
Dr. Vivien CHAN

Financial Statement Frauds

- Financial Statement Fraud
  - Deliberate misrepresentation of the financial condition of a company, e.g. omission of amounts or disclosures in the financial statements, with the intention to deceive or mislead the users of the financial statements
- **Top 10 accounting scandals**
  - Waste management (1998)
  - Enron (2001)
  - WorldCom (2002)
  - Tyco (2002)
  - HealthSouth (2003)
  - Freddie Mac (2003)
  - American International Group (AIG) (2005)
  - Lehman Brothers (2008)
  - Bernie Madoff (2008)
  - Satyam (2009)

Source: https://corporatefinanceinstitute.com/resources/knowledge/other/top-accounting-scandals/

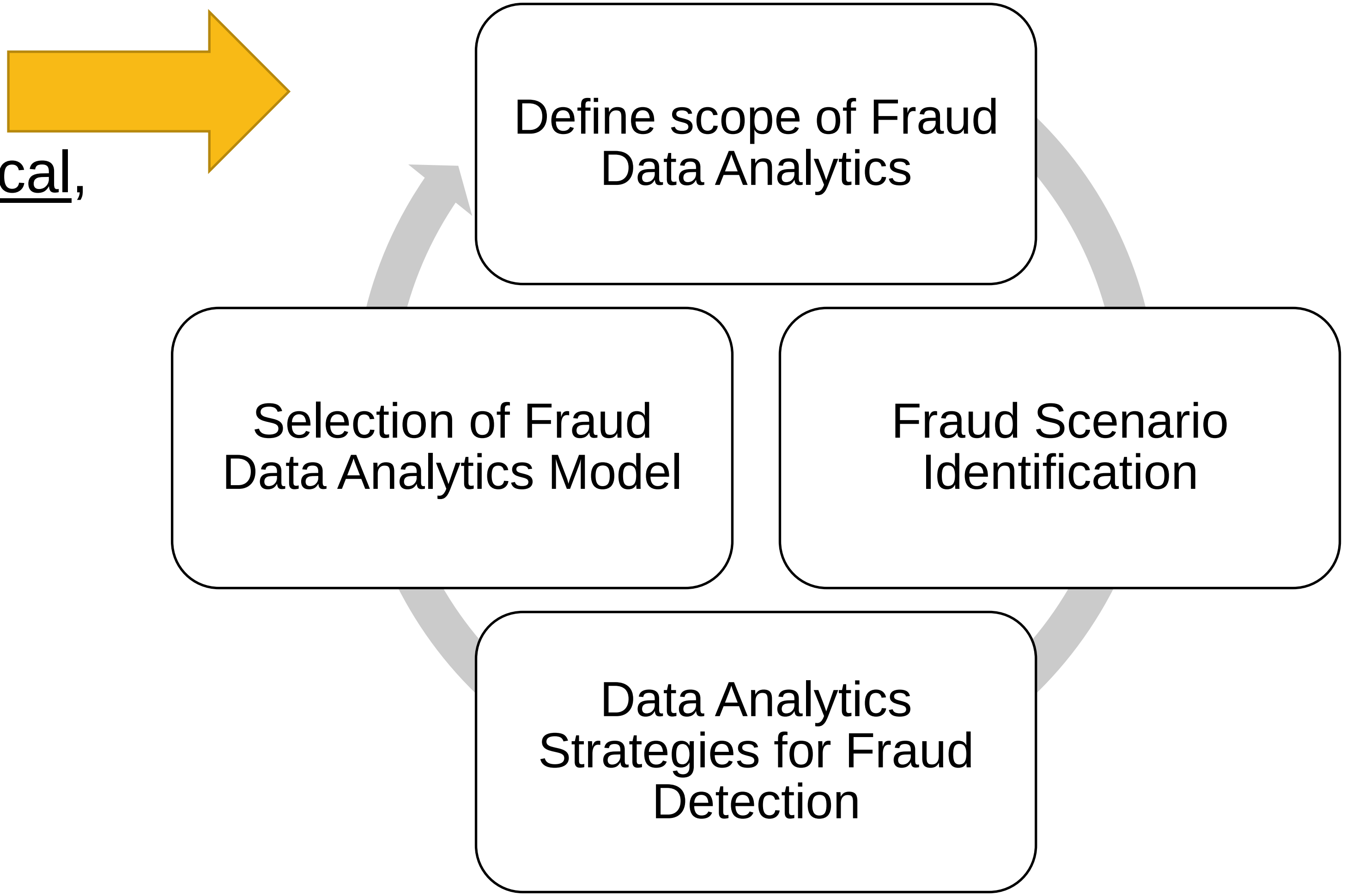
Famous case – Enron Scandal



- Enron deliberately misstated profits, cash flows and understated liabilities with the use of creative, yet questionable accounting methods.
- A disguised loan in 1999 in which the proceeds from the sale of bonds was reported as cash from operations. This overstated operating cash flow by $700 million. With the use of market to market accounting, Enron recognized a very significant amount of future earnings as current income. This allowed a certain business unit to report quarterly profit of $40 million when in fact, this unit was actually operating at a loss. Another loan transaction was understated by $4.85 billion.

(Re-cap) Overview: Fraud Data Analytics Methodology

Staring point.
BUT the process is cyclical,
NOT linear.



Background

- Specific problems of Financial Statement Fraud detections:
  1. the ratio of fraud to nonfraud firms is small
  2. the ratio of false positive to false negative misclassification costs is small
  3. the attributes used to detect fraud are relatively noisy, where similar attribute values can signal both fraudulent and nonfraudulent activities; and
  4. fraudsters actively attempt to conceal the fraud, thereby taking fraud firm attribute values look similar to nonfraud firm attribute values.

Source: Perols, Johan. (2010). Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. Auditing A Journal of Practice & Theory. 30. 10.2308/ajpt-50009.

Background

- Data sample

Panel A: Fraud Firms

Firms investigated by the SEC for fraudulent financial reporting from 4Q 1998 through 4Q 2005	745
Less: Financial companies	(35)
Less: Not annual (10-K) fraud	(116)
Less: Foreign companies	(9)
Less: Not-for-profit organizations	(10)
Less: Registration, 10-KSB, and IPO-related fraud	(78)
Less: Fraud year missing	(13)
Less: Duplicates	(287)
Remaining Fraud Observations	197
Add: Fraud firms from Beasley (1996)	75
Less: Not in Compustat or CompactD for first fraud year or four prior years or I/B/E/S for first fraud year	(221)
Usable Fraud Observations	51

Panel B: Nonfraud Firms

Nonfraud Observations	15,934

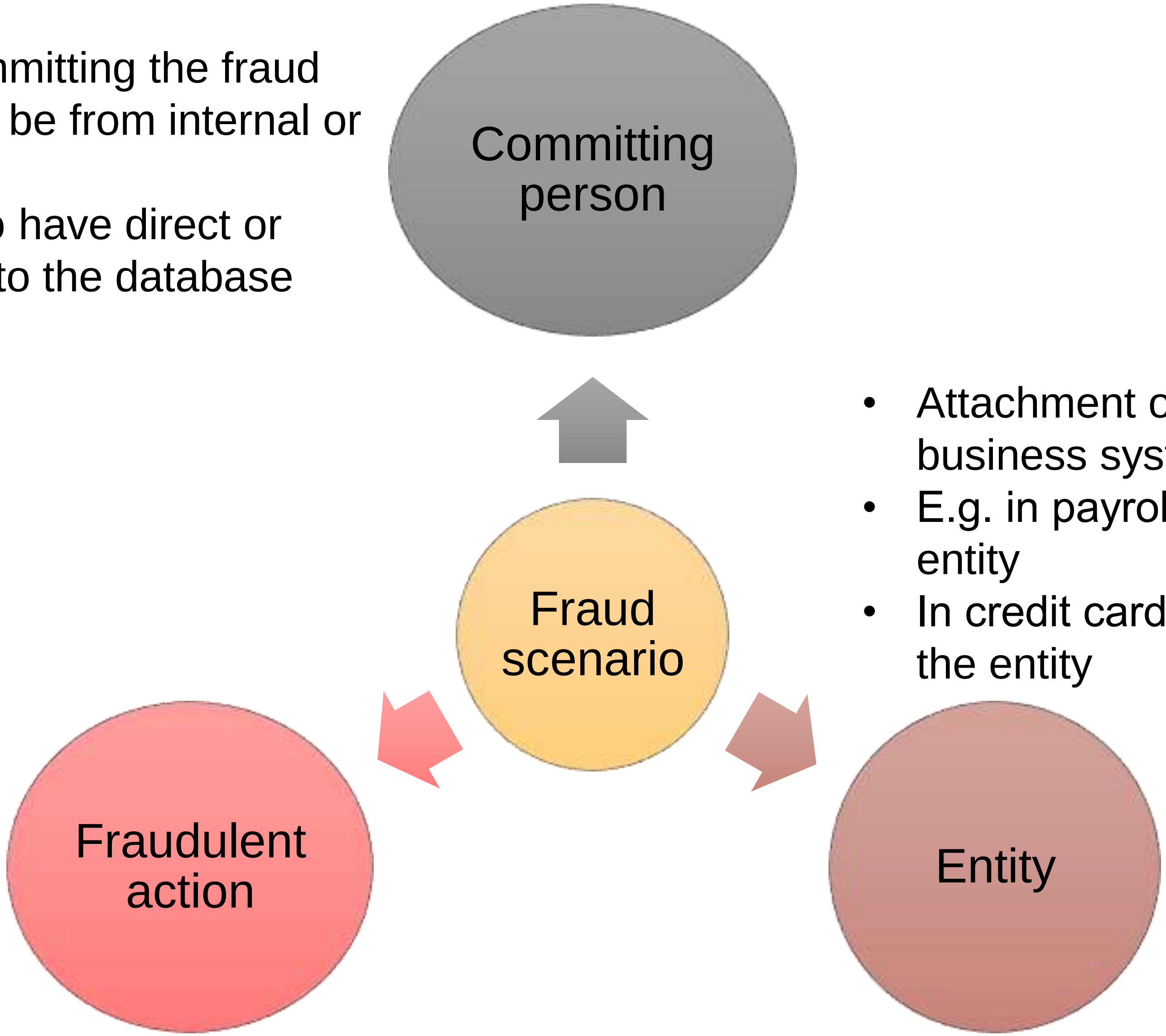
Source: Perols, Johan. (2010). Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. Auditing A Journal of Practice & Theory. 30. 10.2308/ajpt-50009.

1: Define scope of Fraud Data Analytics

- What is the objective and scope of fraud data analytics?
  - To identify the algorithms and predictors to use when creating new models for financial statement fraud detection under specific class and cost imbalance ratios

2: Fraud Scenario Identification (Re-visit)

- The person committing the fraud
- The person can be from internal or external
- The person who have direct or indirect access to the database



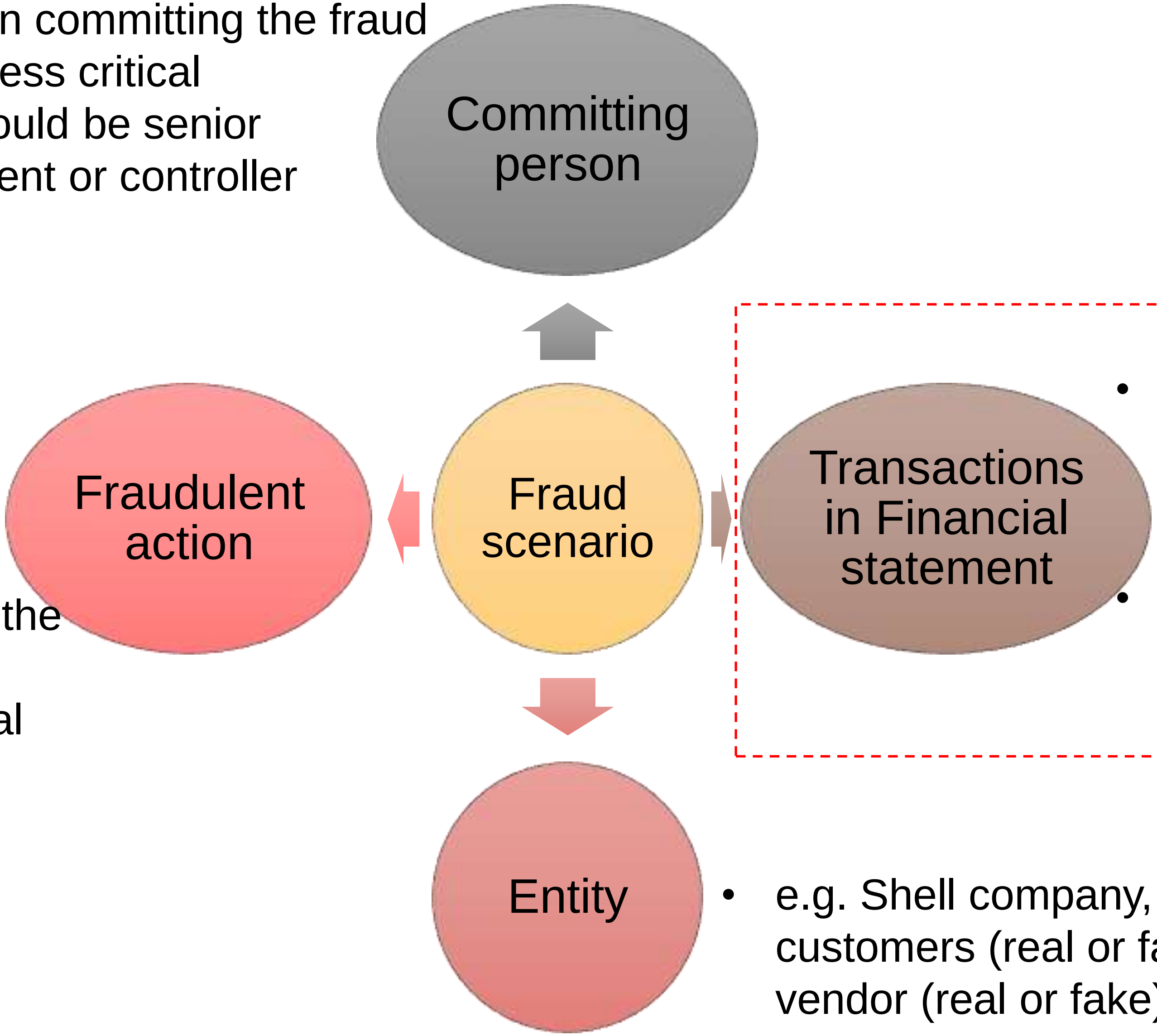
- Attachment of transaction in the business system
- E.g. in payroll system, ‘employee’ is the entity
- In credit card system, ‘card number’ is the entity

- Fraudulent action links committing person and entity
- E.g. payment of vendor without purchase order

Inherent Fraud Scheme for Financial Statement Fraud

- The person committing the fraud would be less critical
- Usually would be senior management or controller

- Fraudulent action describes how the transaction is recorded
- Concerns whether the transaction or the entity is false or real

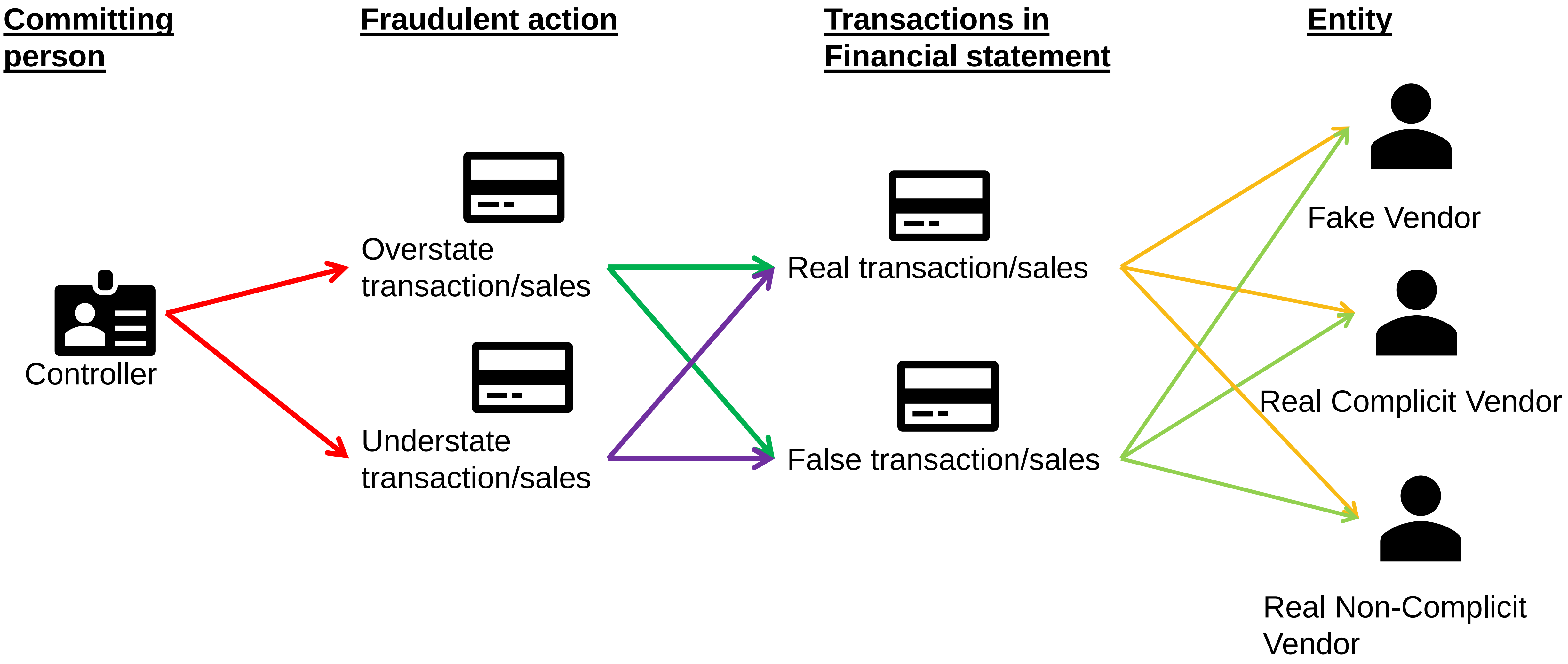


- Direction of misstatement, i.e. general ledger account is overstated or understated; which financial year, etc.
- General ledger account, i.e. transactions recorded is real or fake

- e.g. Shell company, customers (real or fake), vendor (real or fake)

2: Fraud Scenario Identification

- Create the permutation of fraud scenarios



2: Fraud Scenario Identification

Some example predictor attributes:

- number of auditor turnovers
- total discretionary accruals
- Big 4 auditor
- accounts receivable
- allowance for doubtful accounts
- accounts receivable to total assets
- accounts receivable to sales
- whether meeting or beating forecast
- evidence of CEO change
- sales to total assets
- inventory to sales
- unexpected employee productivity
- percentage of executives on the board of directors
- whether accounts receivable grew by more than 10 percent
- allowance for doubtful accounts to net sales
- current minus prior year inventory to sales
- gross margin to net sales
- evidence of CFO change
- holding period return in the violation period
- property plant and equipment to total assets
- value of issued securities to market value
- fixed assets to total assets;
- days in receivables index
- industry ROE minus firm ROE
- positive accruals dummy
- whether gross margin grew by more than 10 percent
- allowance for doubtful accounts to accounts receivable
- total debt to total assets

Source: Perols, Johan. (2010). Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. Auditing A Journal of Practice & Theory. 30. 10.2308/ajpt-50009.

2: Fraud Scenario Identification

- Example techniques used to overstate an asset:
  - Recording an asset that does not exist
  - Recording a real asset before the liability occurs
  - Recording a real asset that is not owned by the company
  - Improper capitalization of a false expense
  - Improper capitalization of a real expense
  - Reporting the asset in the wrong section of the balance sheet
- Example techniques used to understate an asset:
  - Failure to record a real asset
  - Failure to capitalize a real expense
  - Failure to record an asset in the proper period
  - Reporting the asset in the wrong section of the balance sheet

3: Data Analytics Strategies for Fraud Detection

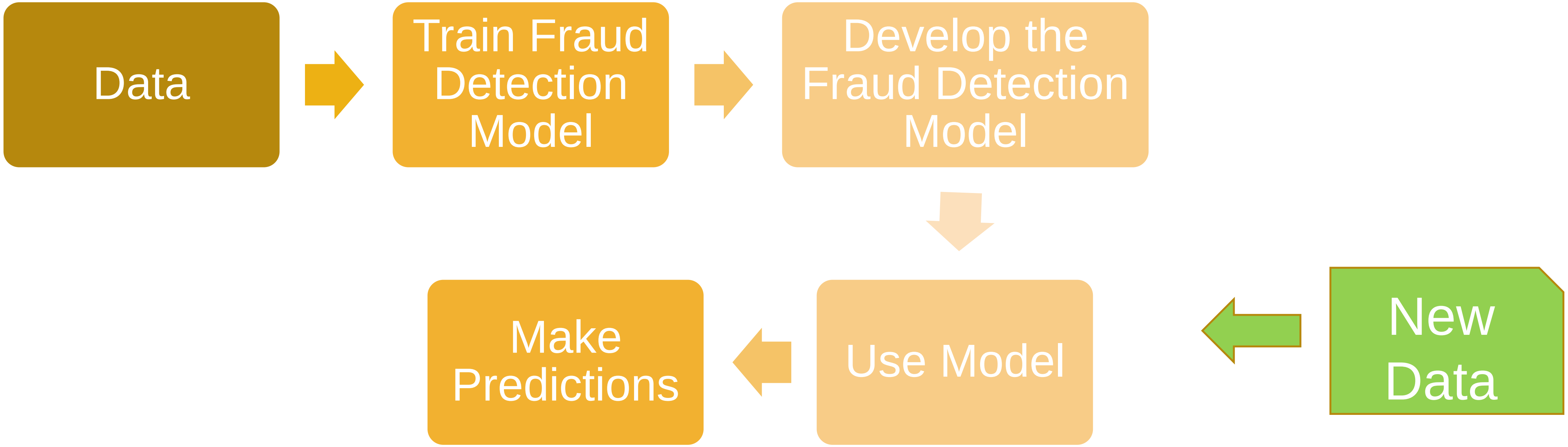
Fraud Analytics	Predictive Analytics (Modeling)
Uses historical data to detect fraud that has already occurred	Uses historical data to predict future outcomes
Linear process; the steps are performed in order, and typically the process is not repeated	Nonlinear process; steps can be skipped, and the process is reiterative
A hypothesis is formed at the beginning of the fraud engagement	Models are defined and created based on the particular business process
Analysis stage may continue longer than expected if additional hypotheses are formed	Process is repeated if new data or different variables are discovered
Hypothesis is tested and amended as necessary	Models are tested to determine success; modifications are made as necessary
Fraud analysis is used to locate fraud and can provide a model for future detection	Predictive modeling is used to complement the fraud analysis by creating a process to show red flags

3: Data Analytics Strategies for Fraud Detection

Fraud Analytics	Predictive Analytics (Modeling)
Data quality is important to the analyst's ability to discover the fraud	Data quality is important to the success of the model
Uses all available data	Uses a sample of the available data
Constructs data (mean, median, mode) for statistical analysis purposes	Constructs data to fill in missing variables
Fraud analysis is performed as needed, not on a regular recurring basis, and ends with a final conclusion	Models are repetitive and cyclical in nature; they are always in process
Looks for anomalies in the data	Looks for anomalies in the data
Outcome cannot be predicted and is known only after the dissemination stage	Outcome or final goal must be specifically defined

4: Selection of Fraud Data Analytics Model

Fraud Detection Model



Frequency of re-training the model depends on:

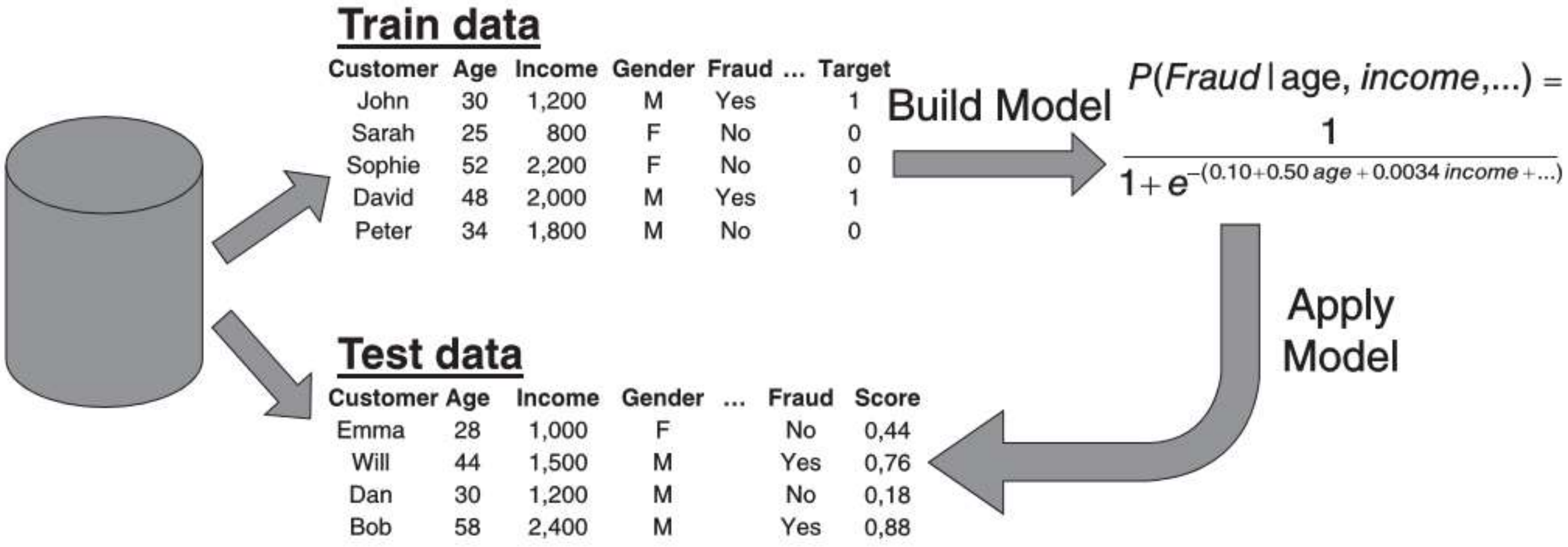
- Volatility of the fraud behaviour
- Detection power of the current model
- Amount of (similar) confirmed cases already available in the database
- Rate at which new cases are being confirmed
- Required effort to retrain the model

03 IMBALANCE DATA HANDLING

Dr. Vivien CHAN

Sample Data Set

- Split the sample data set into 2-3 datasets



Train data

Customer	Age	Income	Gender	Fraud	...	Target
John	30	1,200	M	Yes		1
Sarah	25	800	F	No		0
Sophie	52	2,200	F	No		0
David	48	2,000	M	Yes		1
Peter	34	1,800	M	No		0

$$P(\mathit{Fraud} \mid \text{age}, \mathit{income}, \ldots) = \frac{1}{1+e^{-(0.10+0.50\,\mathit{age}+0.0034\,\mathit{income}+\ldots)}}$$

Test data

Customer	Age	Income	Gender	...	Fraud	Score
Emma	28	1,000	F		No	0,44
Will	44	1,500	M		Yes	0,76
Dan	30	1,200	M		No	0,18
Bob	58	2,400	M		Yes	0,88

Figure 4.34 Training Versus Test Sample Set Up for Performance Estimation

Source: Fraud Analytics using Descriptive, Predictive, and Social Network Techniques, 1st ed, John Wiley & Sons Inc.

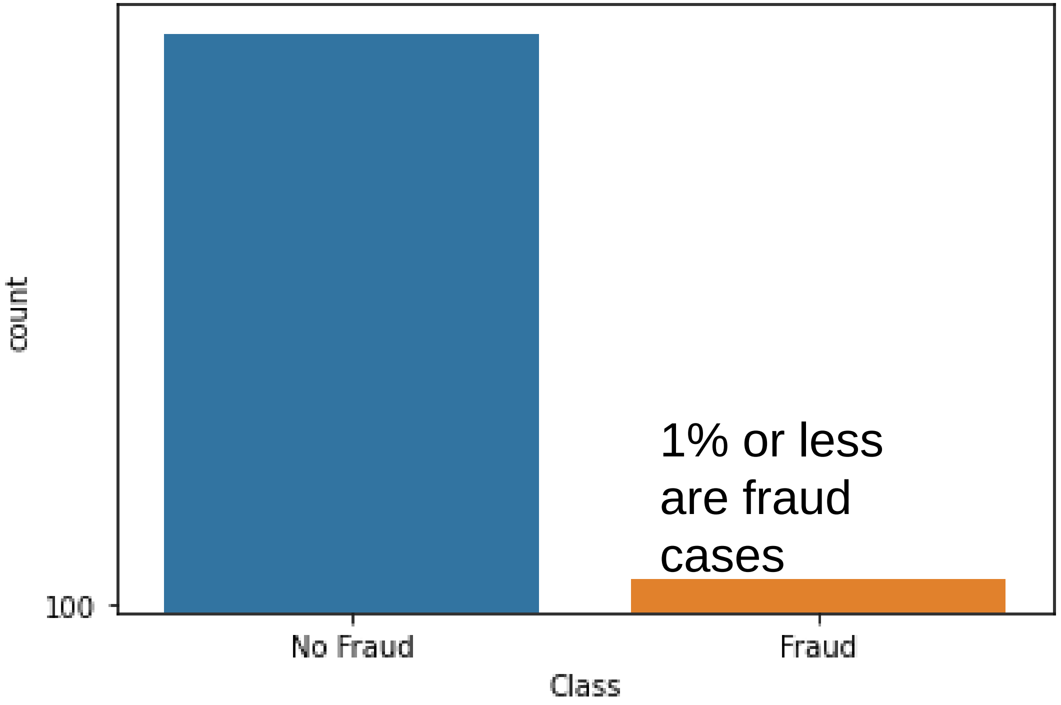
Splitting the data set

- Observations used for training should not be used for testing or validation
- If validation data set is not required,
  - 70% for training
  - 30% for testing
- If validation data set is required,
  - 40% for training data
  - 30% for validation data
  - 30% for testing data

Source: Fraud Analytics using Descriptive, Predictive, and Social Network Techniques, 1st ed, John Wiley & Sons Inc.

What is imbalance dataset?

- Imbalance dataset also known as skewed dataset
- Imbalanced datasets are a special case for classification problem where the class distribution is not uniform among the classes.
- Typically, they are composed by two classes: The majority class and the minority class.

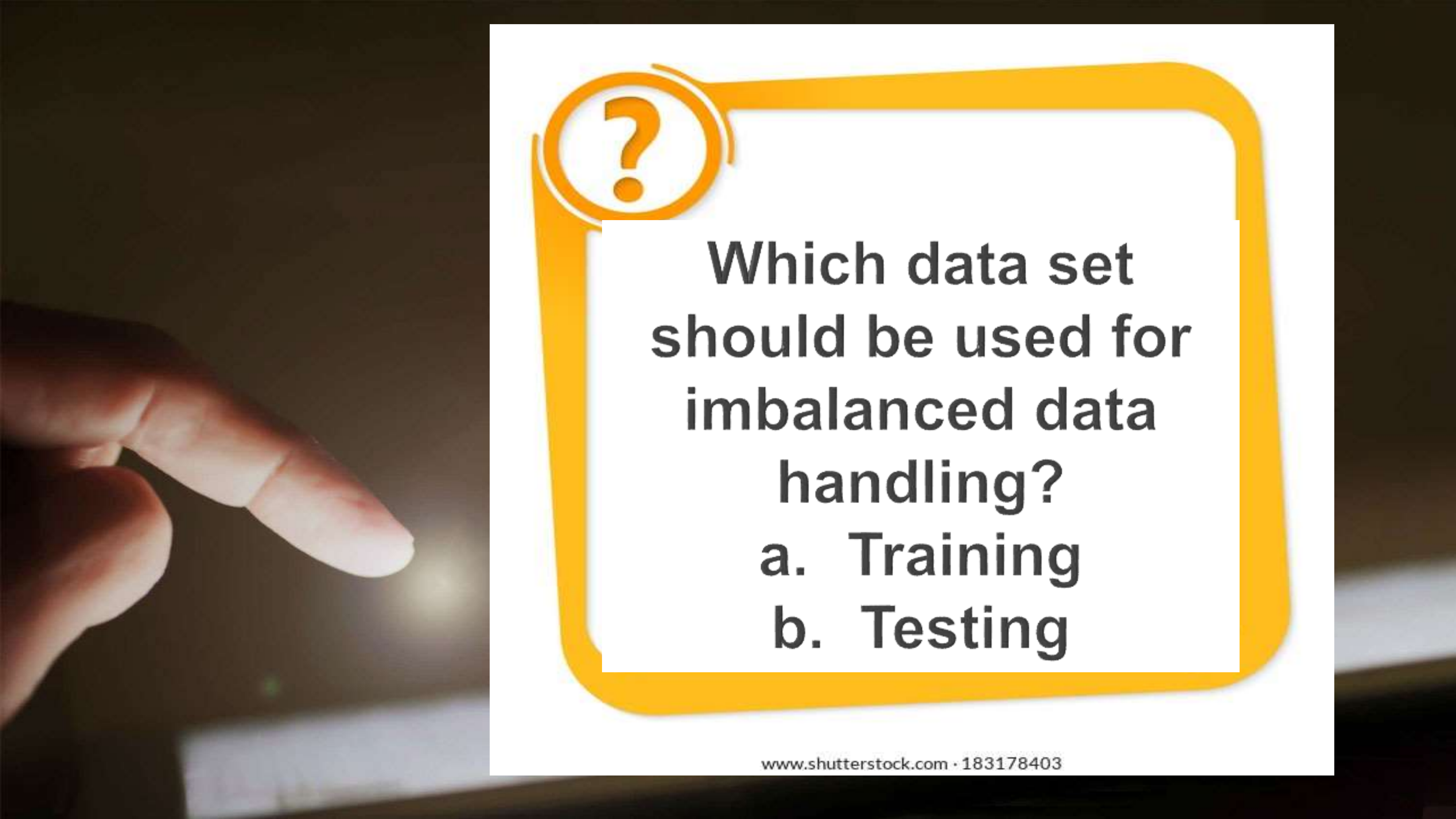


What are the problems of imbalance dataset?

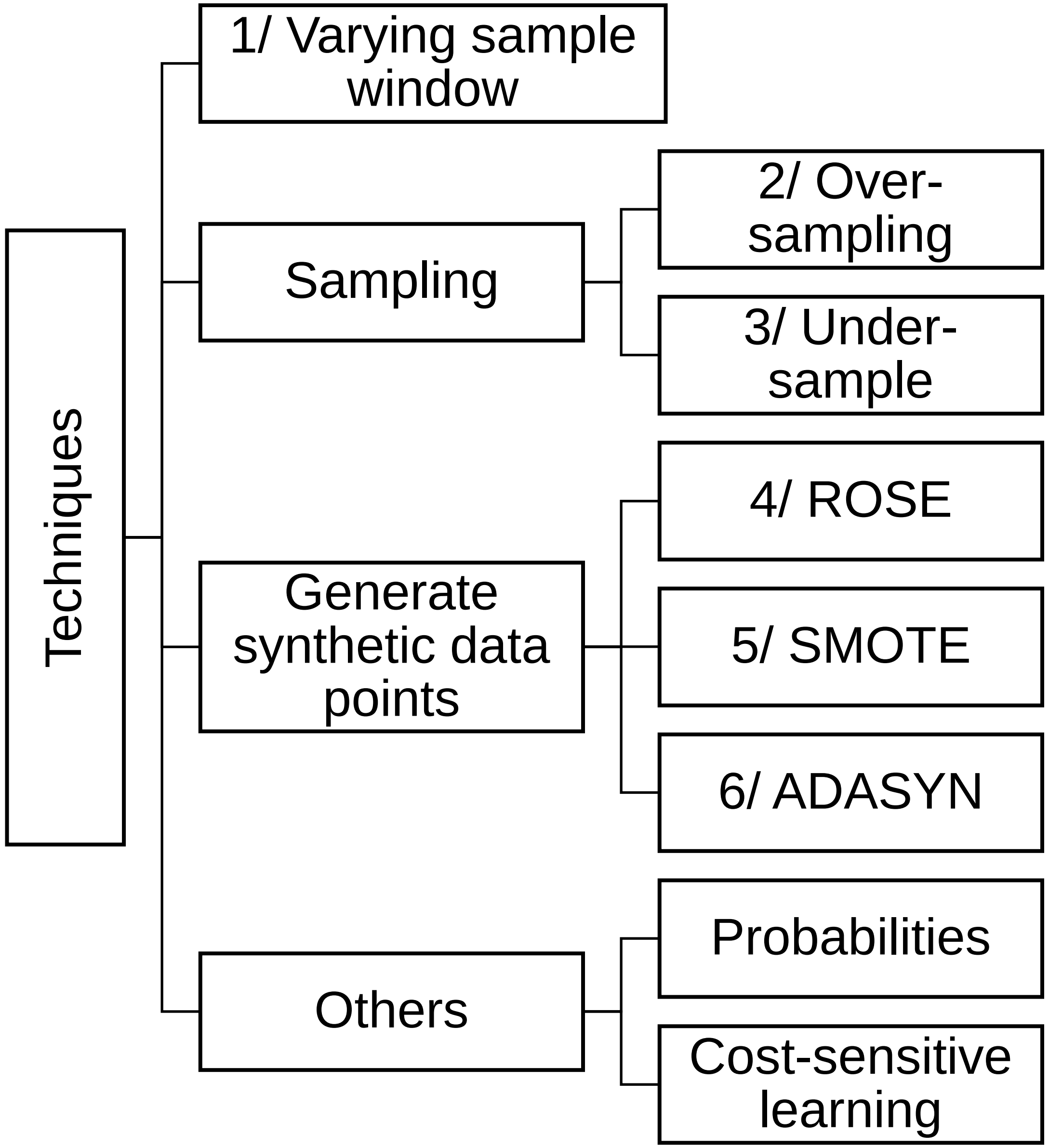
What are the problems?

- Problems of imbalance dataset in machine learning:
  - Most machine learning models assume an equal distribution of classes
  - A model may focus on learning the characteristics of majority class due to the abundance of samples available for learning
  - Many machine learning models will show bias towards majority class, leading to incorrect conclusions

- Slight imbalance vs Severe imbalance
  - If the data set is only slightly imbalance (e.g. ratio of 4:6), can still be used for training


Which data set should be used for imbalanced data handling?
a. Training
b. Testing
www.shutterstock.com · 183178403

How to handle imbalanced dataset



How to handle imbalanced dataset

- 1/ Varying the Sample Window
  - Increase the number of fraudsters by increasing the time horizon
    - e.g. instead of taking samples from 6 months, use a 12-month window
  - Sample every fraudsters twice or more as shown in fig 4.47
    - e.g. varying the timeline of observation period to obtain additional set of sample data which are similar but not exactly the same

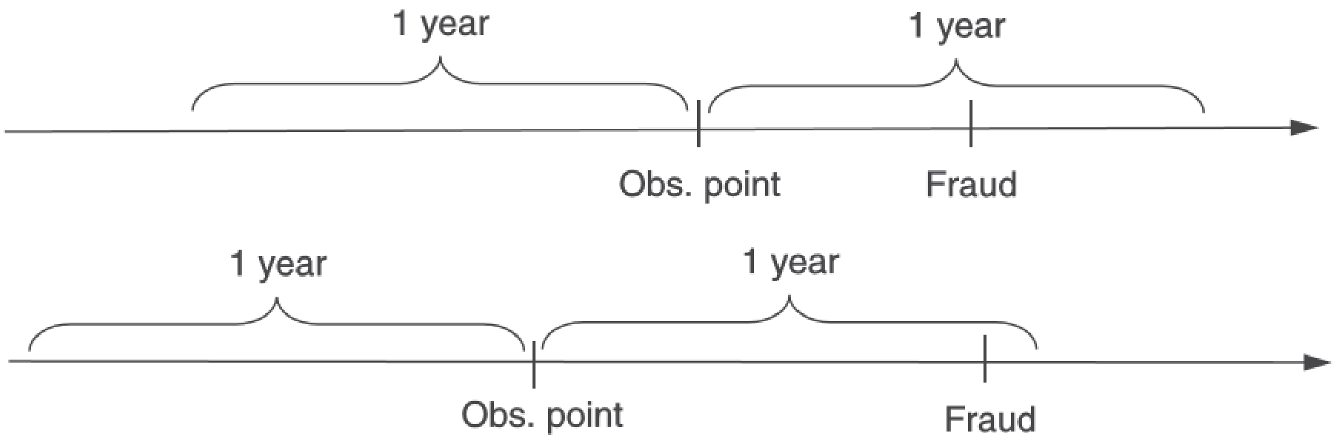
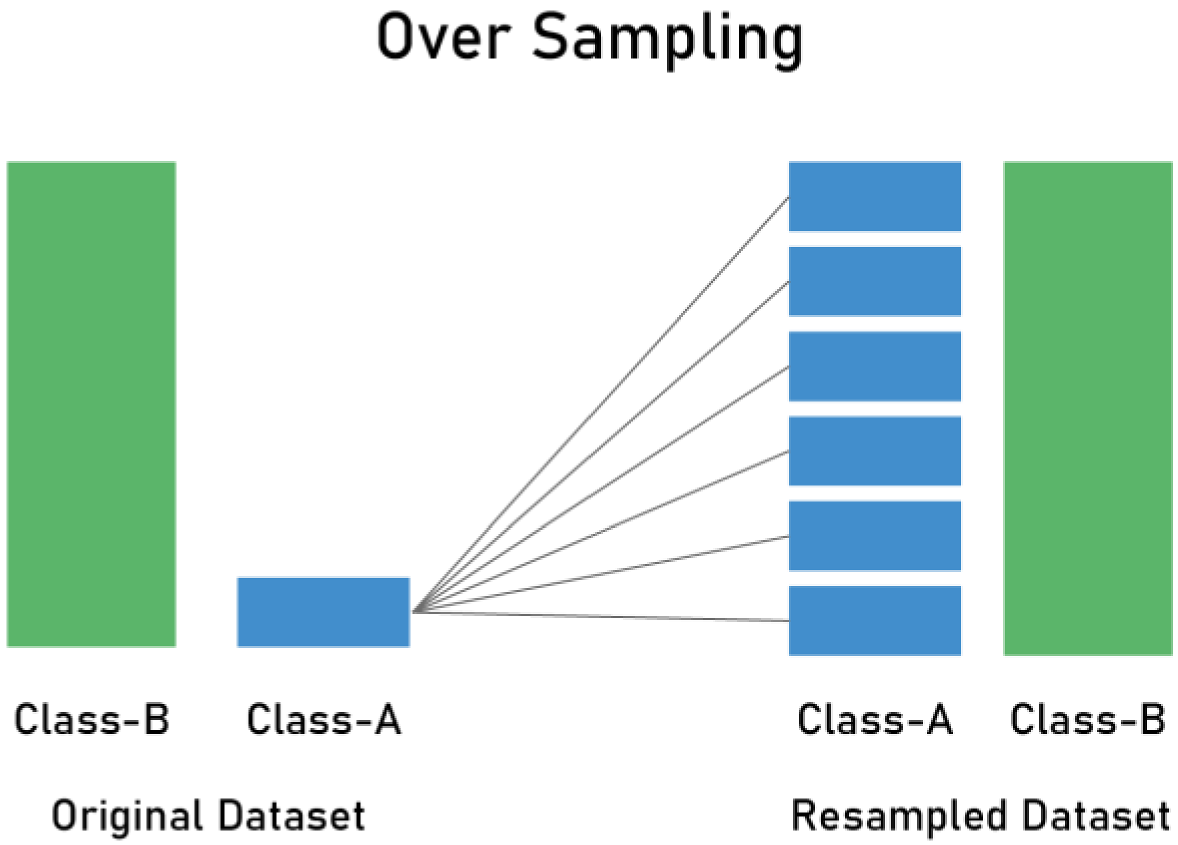


Figure 4.47 Varying the Time Window to Deal with Skewed Data Sets

How to handle imbalanced dataset



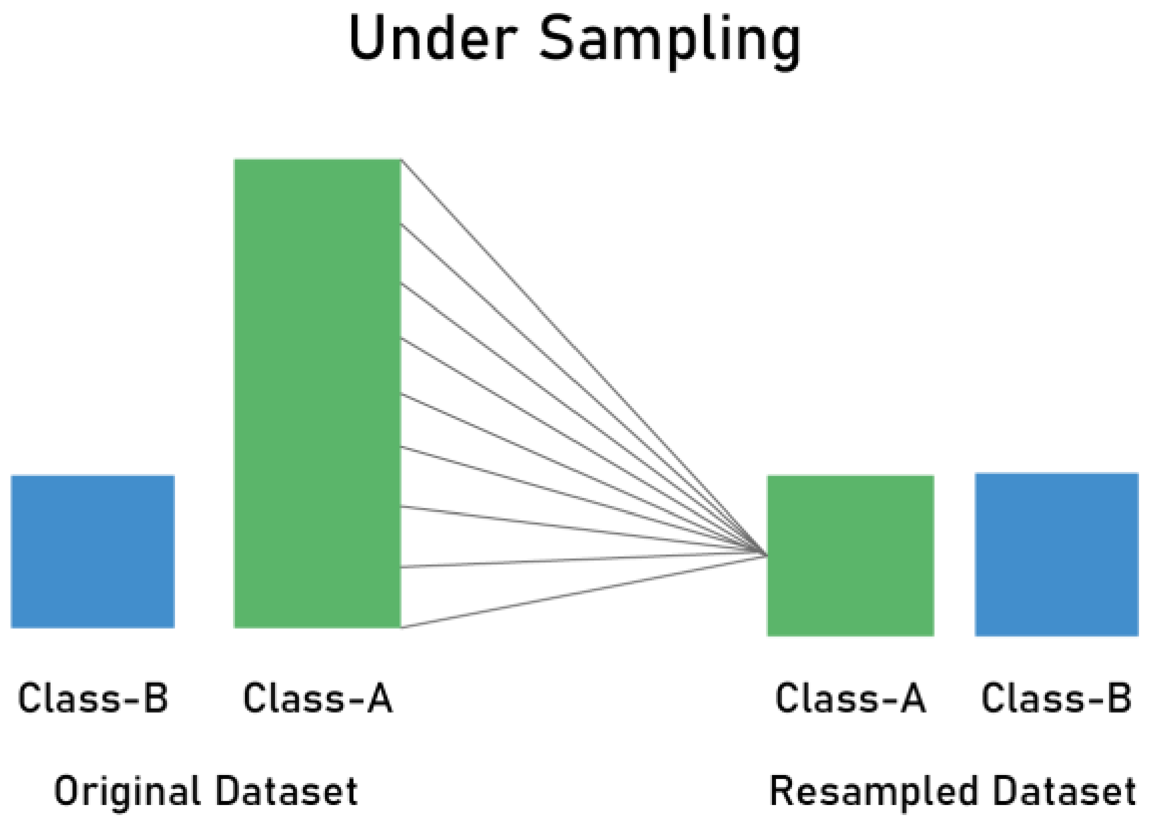
2/ Over sampling

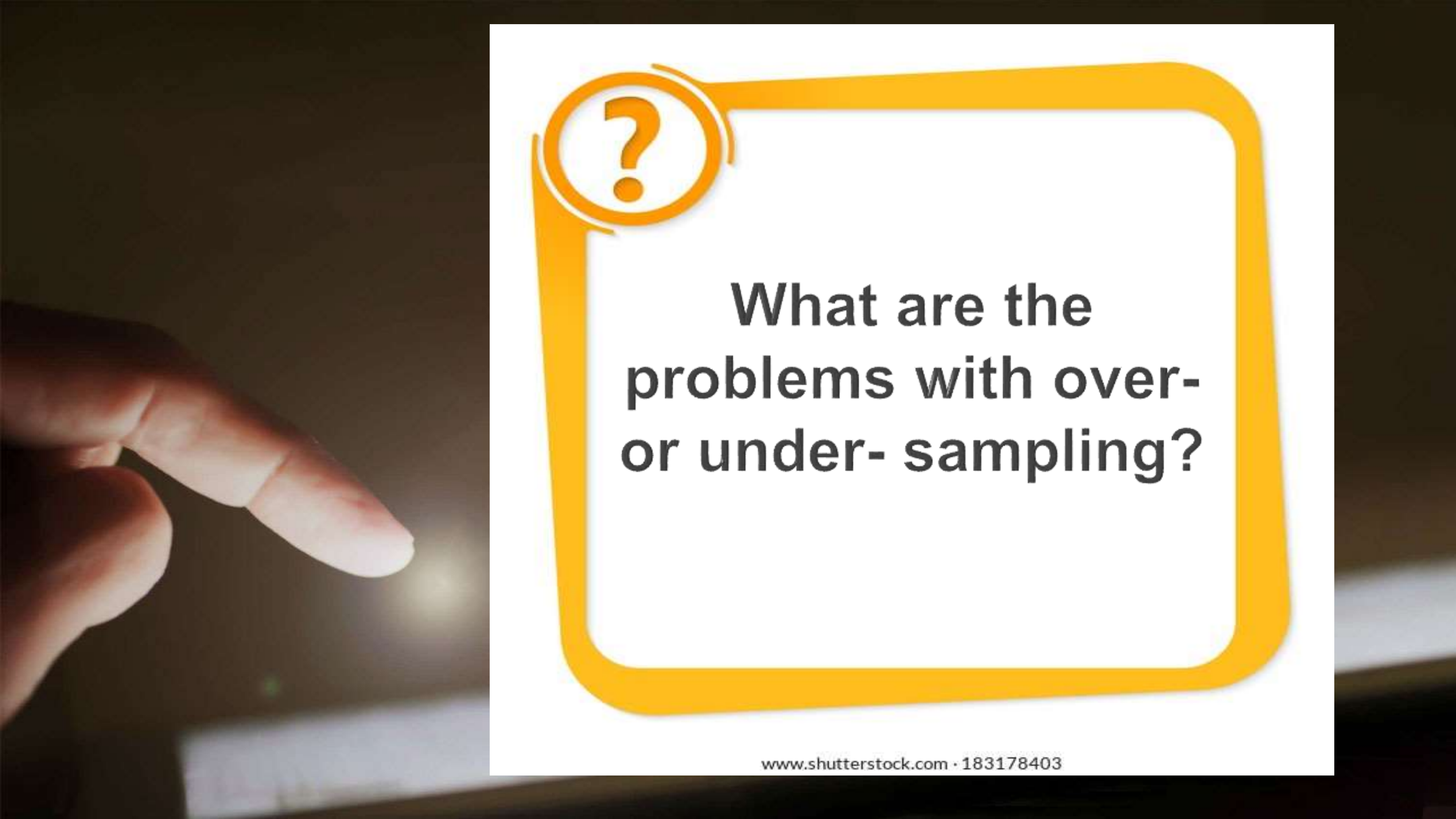
- It helps to increase the number of minority class examples in the dataset.
- One of the main advantages of oversampling is no information is lost from both the majority and minority classes during the process.

How to handle imbalanced dataset

3/ Under sampling

- it helps to reduce the number of majority class examples in the dataset.




What are the problems with over- or under- sampling?
www.shutterstock.com · 183178403

How to handle imbalanced dataset

4/ ROSE (Random Over Sampling Example) (Menardi and Torelli, 2014))

- combines techniques of oversampling and undersampling by generating an augmented sample of data (especially belonging to the rare class)
- thus helping the classifier in estimating a more accurate classification rule, because the same attention will be addressed to both the classes
- the synthetic generation of new examples allows for strengthening the process of learning as well as estimating the distribution of the chosen measure of accuracy

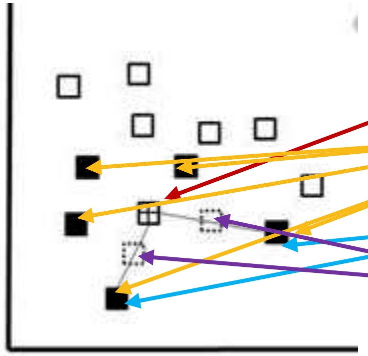
How to handle imbalanced dataset

5/ SMOTE (Synthetic Minority Over-Sampling Technique) (Chawla et al., 2001)

- Creates synthetic observations based upon the existing minority observations
- Combines the synthetic oversampling of the minority class with undersampling the majority class
- SMOTE proven to be better than either under-/over-sampling. It is also proven to be valuable for fraud detection

How to handle imbalanced dataset

5/ SMOTE (Synthetic Minority Over-Sampling Technique) (Chawla et al., 2001)



Steps

1/ Select one minority class observation

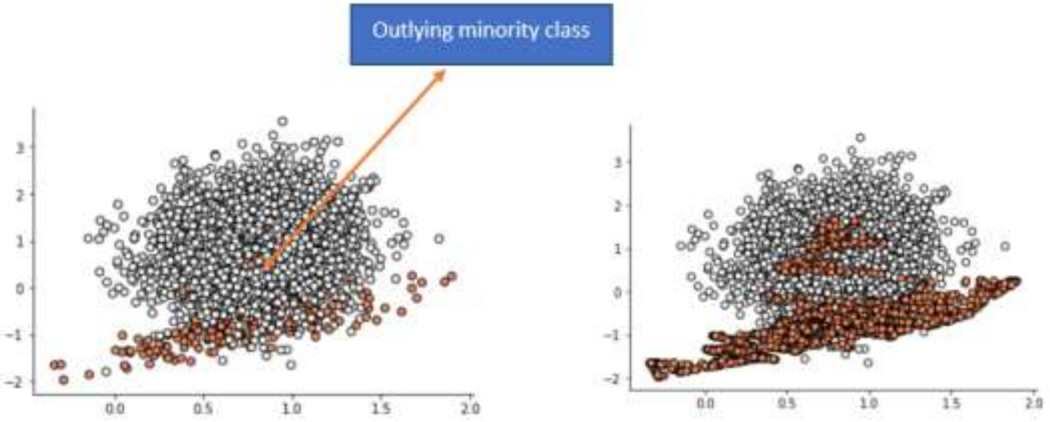
2/ Calculate 5 nearest neighbour

3/ Randomly select 2 nearest neighbour

4/ Randomly create 2 new observations as samples

PROBLEM with SMOTE

If there are observations in the minority class which are outlying and appears in the majority class, it causes a problem for SMOTE, by creating a line bridge with the majority class.



How to handle imbalanced dataset

6/ ADASYN (Adaptive Synthetic sampling)

- a generalization of the SMOTE algorithm
- it takes into account the distribution of density
- it measures the K-nearest neighbors for all minority instances, then calculates the class ratio of the minority and majority instances to create new samples
- For example,
  - Impurity ratio is calculated for all minority data points
  - Higher the ratio, more synthetic data points are created. E.g. the synthetic data points of Obs3 will be 4 times that of Obs2

Example: k=5, i.e. only look at 5 neighbours

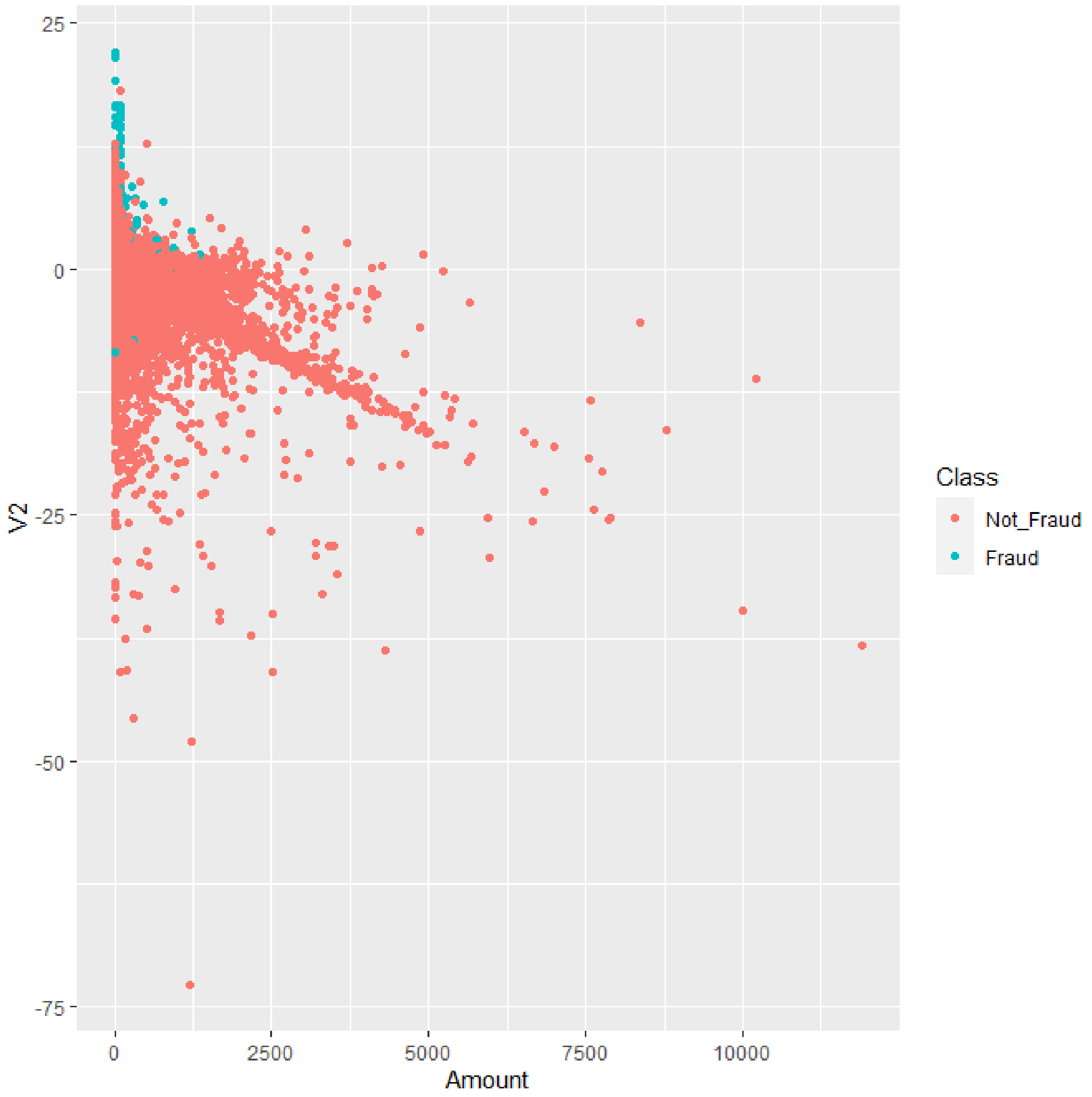
Fraud class data points	**Fraud class Neighbours**	**Non-fraud class Neighbours**	**Impurity Ratio**
Observation 1	3	2	0.4
Observation 2	4	1	0.2
Observation 3	1	4	0.8
Observation 4	5	0	0

Packages in R

1- ROSE: *The package only implements the algorithm Random Over Sampling*
Link: *ROSE*
https://cran.r-project.org/web/packages/ROSE/index.html

2- DMwR: *The package reads as "Data Mining with R" and comes with implementation of SMOTE algorithm. SMOTE algorithm uses nearest neighbor concept to oversample the minority class.*
Link: *DMwR*
https://cran.r-project.org/src/contrib/Archive/DMwR/

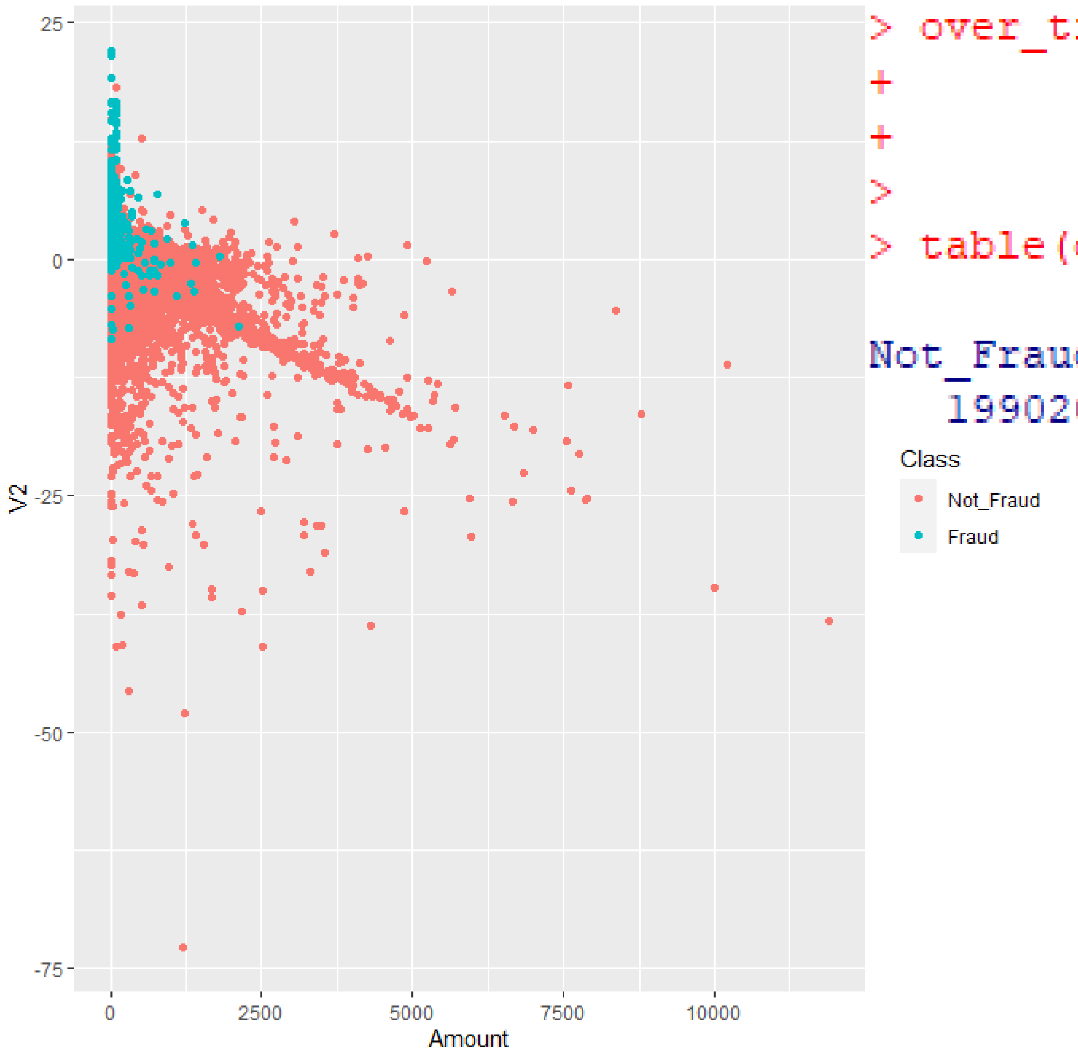
Example – Original Dataset




```
> # class ratio initially
> table(train$Class)

Not_Fraud      Fraud
   199020        344
```

Example - Oversampling



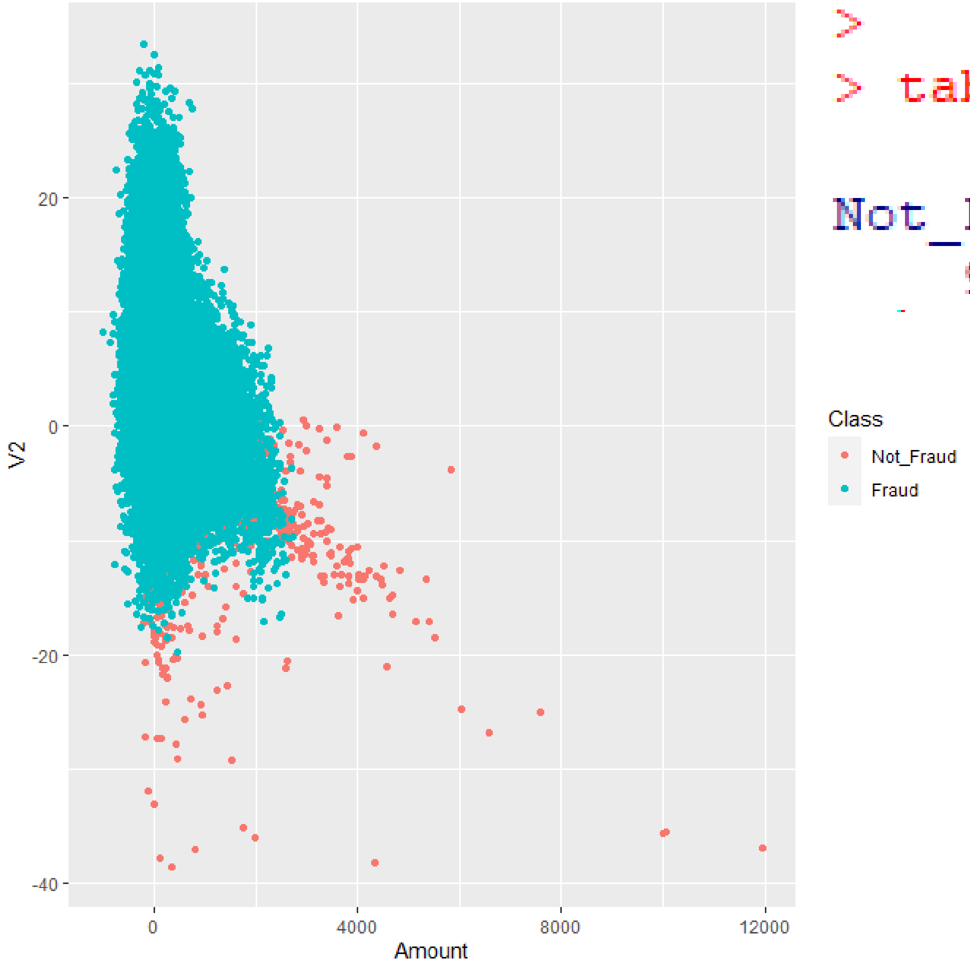

```
> # oversampling
> set.seed(9560)
> over_train <- ovun.sample(Class~., data=train,
+                                   p=0.5,
+                                   seed=1, method="over")$data
>
> table(over_train$Class)

Not_Fraud     Fraud
   199020    199283
```

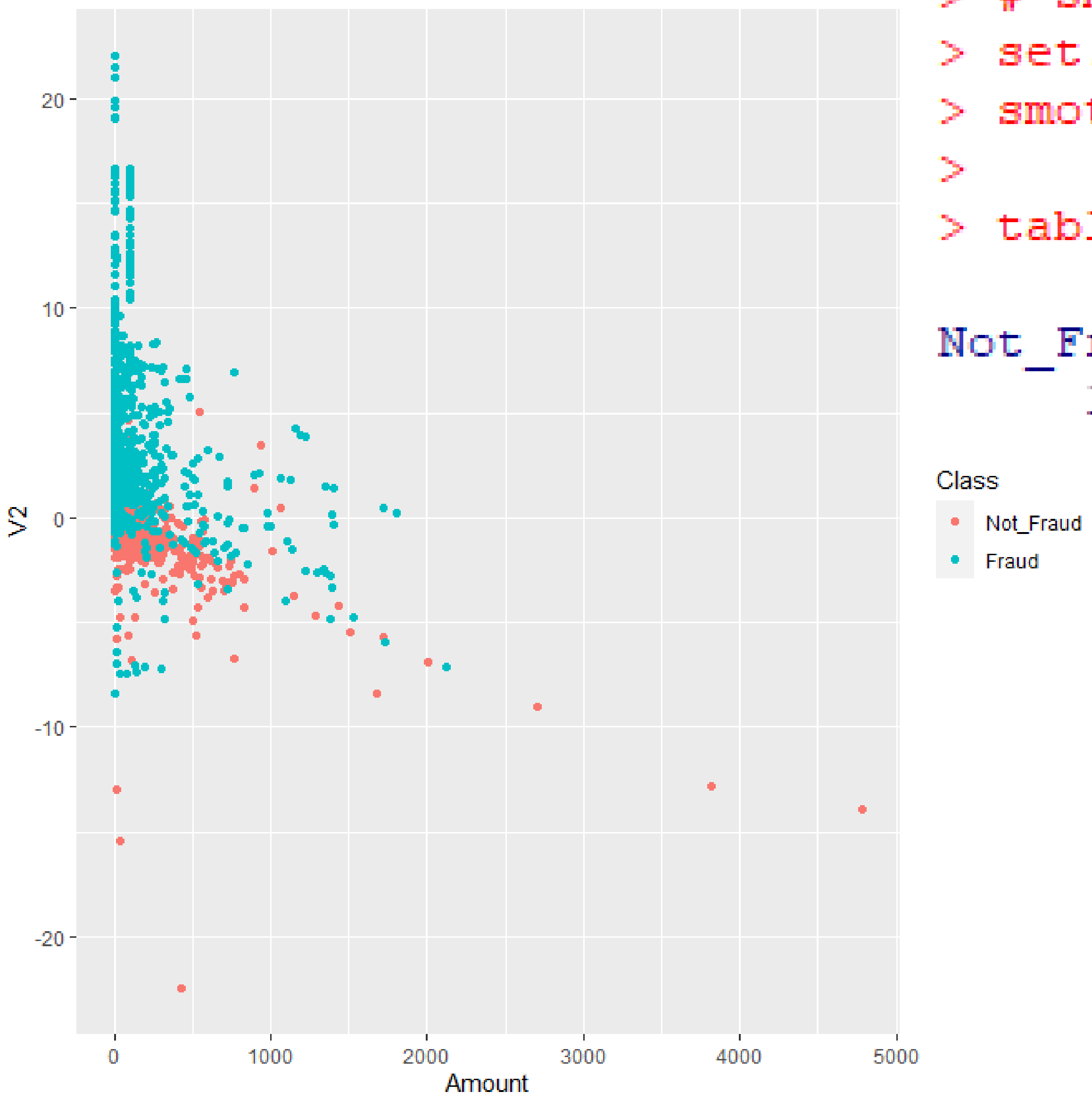
Example - ROSE

```
> # rose
> set.seed(9560)
> rose_train <- ROSE(Class ~ ., data  = train)$data
>
> table(rose_train$Class)

Not_Fraud     Fraud
    99844     99520
```



Example - SMOTE




```
> # smote
> set.seed(9560)
> smote_train <- SMOTE(Class ~ ., data  = train)
>
> table(smote_train$Class)

Not_Fraud     Fraud
     1376      1032
```

04
Linear and Logistic Regression

Dr. Vivien CHAN

Linear Regression

- Most commonly used technique to model a continuous target variable
- For example, Car insurance fraud detection
  - a linear regression model can be defined to model the amount of fraud in terms of the age of the claimant, claimed amount, severity of accident, etc.

$$Amount\ of\ fraud = \beta_0 + \beta_1 Age + \beta_2 ClaimedAmount + \beta_3 Severity + \ \ldots$$

- General formulation of Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \ \ldots \ + \beta_N X_N$$

Y = target variable (or dependent variable)

$X_{1, \ldots} X_N$ = explanatory variables (or independent variables)

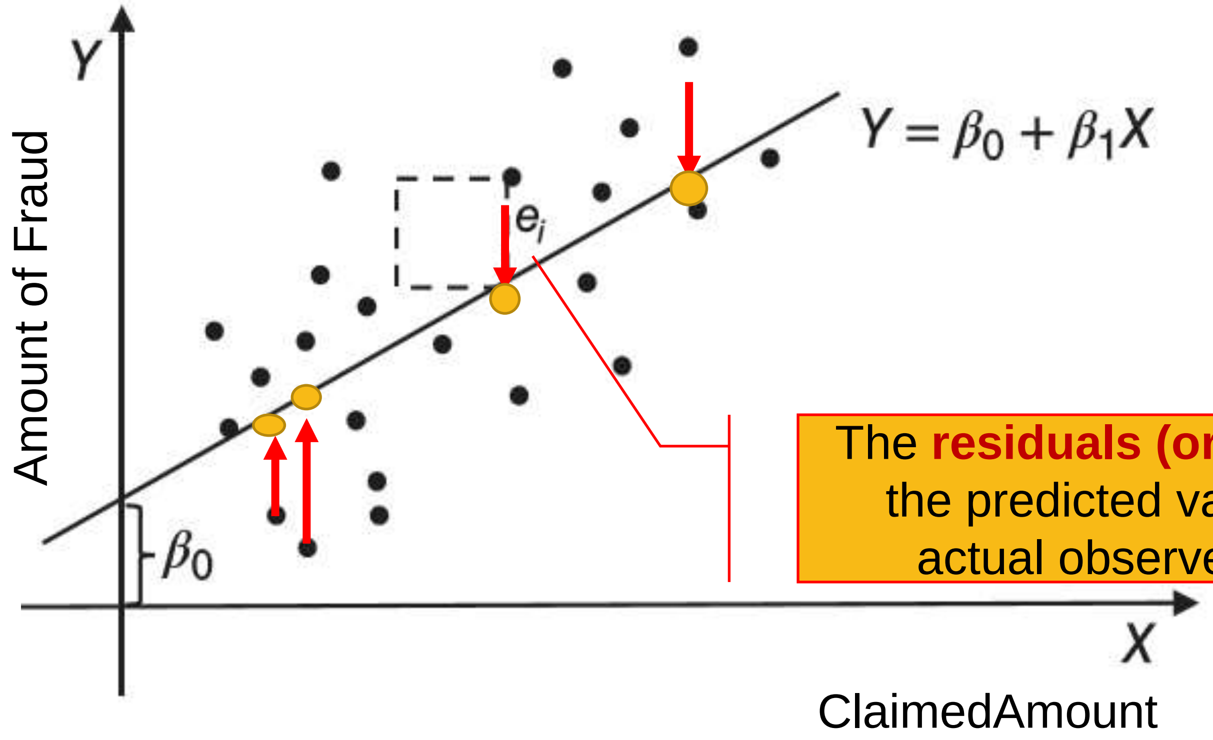
β = parameters measuring the impact on the target variable Y of each of the individual explanatory variables $X_{1, \ldots} X_N$

Linear Regression

- Question: How to find the best fit straight line through the data?
- Ans: By minimizing the sum of all error squares (MSE = mean square error)

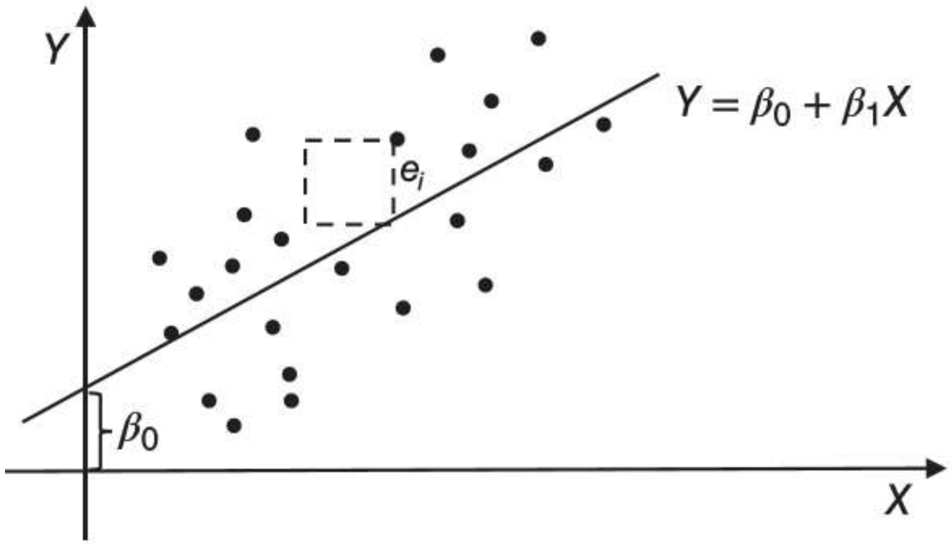
$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error
n = number of data points
Y_i = observed values
$\hat{Y}_i$ = predicted values



The **residuals (or error terms)** are the predicted values minus the actual observed values of Y

How to interpret Linear Regression output?



Ordinary Least Square (OLS) Regression

$$Y = \beta_0 + \beta_1 X_1 + \ \dots \ + \beta_N X_N$$

Example:

$$Amount\ of\ fraud = \beta_0 + \beta_1 Age + \beta_2 ClaimedAmount + \beta_3 Severity + \ \dots$$

Slope = Positive or negative relation between X (e.g. Age, ClaimedAmount, Severity) and Y (e.g. Amount of fraud)

$\beta_{1\dots}\beta_N$ = Regression Coefficient of a variable i.e. the change in the response based on 1-unit change in the corresponding explanatory variable, keeping all other variables held constant.

β_0 = Intercept coefficient i.e. expected mean value of Y when all X=0. However, if X never = 0, Y will have no meaning.

How to evaluate a Linear Regression model?

Dr. Vivien CHAN

Performance Measures for regression models

- For classification models,
  - the output is categorical data
  - the measure of the performance is counting the % of correctly predicted value.
- For regression models,
  - the output is a continuous number
  - the measure of the performance is how “close” the predicted value is to the actual value.
  - i.e. What is the “loss” incurred by the model in predicting the actual value of a data point?
  - Or, any deviation from the actual value is an error

Error = Y (actual) — Y (predicted)

Performance Measures for regression models

- Commonly used performance measures:
  - Mean Absolute Error (MAE)
  - Mean Squared Error (MSE)
  - Root Mean Squared Error (RMSE)
  - R-Squared
  - Adjusted R-squared

MAE, MSE, RMSE

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

- MAE
  - where $\mathbf{y_i}$ is the actual expected output and $\mathbf{\hat{y}_i}$ is the model's prediction.
  - It is the simplest evaluation metric for a regression scenario and is not much popular compared to the other metrics.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

- MSE
  - The error term is squared and thus **more sensitive to outliers** as compared to Mean Absolute Error (MAE).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

- RMSE
  - Since MSE includes squared error terms, we take the square root of the MSE, which gives rise to Root Mean Squared Error (RMSE).

R-squared

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

- R-squared is calculated by dividing the sum of squares of residuals **(SSres)** from the regression model by the total sum of squares **(SStot)** of errors from the average model and then subtract it from 1.
- R-squared is also known as the **Coefficient of Determination**. It *explains the degree to which the input variables explain the variation of the output / predicted variable.*
- The metric helps us to compare our current model with a constant baseline value (i.e. mean) and tells us how much our model is better

Adjusted R-squared

$$\text{Adjusted } R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

- Here, **N**- total sample size (number of rows) and **p**- number of predictors (number of columns)
- The **limitation of R-squared** is that it will either stay the same or increases with the addition of more variables, even if they do not have any relationship with the output variables.
- To overcome this limitation, Adjusted R-square comes into the picture as it penalizes you for adding the variables which do not improve your existing model.
- Hence, if you are building Linear regression on multiple variables, it is always suggested that you use Adjusted R-squared to judge the goodness of the model.
- If there exists only one input variable, R-square and Adjusted R squared are same.

How high R-squared needs to be?

a. As high as possible

b. Low R^2 is better

Other Performance Measures for regression models

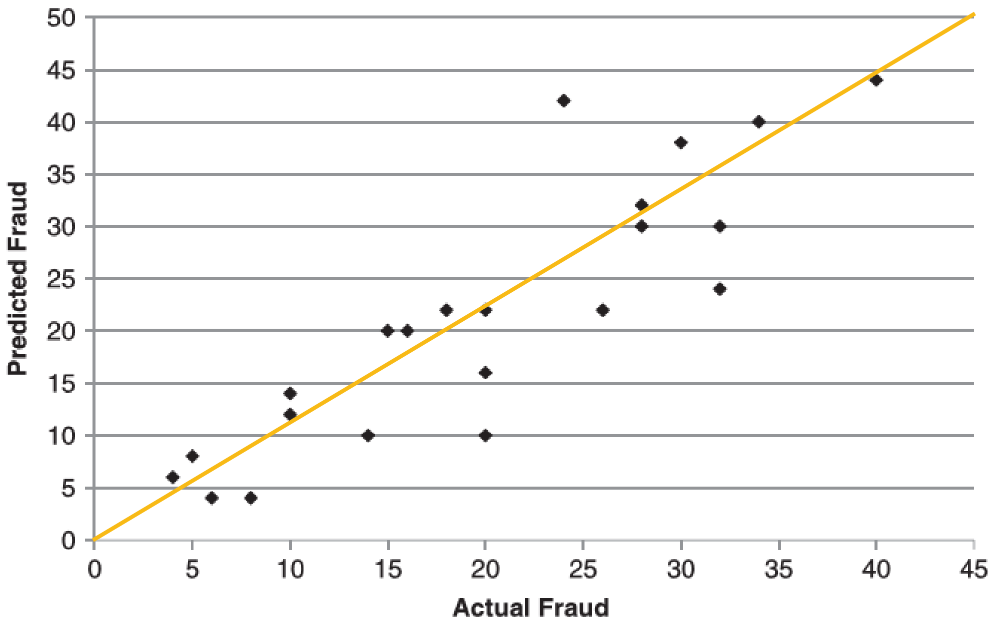


Figure 4.44 Scatter Plot: Predicted Fraud Versus Actual Fraud

- Scatter Plot
  - The more the plot approximates a straight, the better the performance of the regression model

- Pearson Correlation Coefficient

$$corr(\hat{y}, y) = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $\hat{y}_i$ represents the predicted value for observation i, $\hat{y}$ the average of the predicted values, y_i the actual value for observation i, and $\bar{y}$ the average of the actual values. The Pearson correlation always varies between –1 and +1. Values closer to +1 indicate better agreement and thus better fit between the predicted and actual values of the target variable.

Source: Fraud Analytics using Descriptive, Predictive, and Social Network Techniques, 1st ed, John Wiley & Sons Inc.

Linear Regression

Advantages

- Performs exceptionally well for linearly separable data
- Operationally efficient and easy to interpret & implement
- Extrapolation beyond a specific data set

Disadvantages

- Target and exploratory variables must be of linear relation
- Prone to noise and overfitting
- Sensitive to outliers
- Assumes exploratory variables are independent. Might have problem of multicollinearity

Logistic Regression

- Linear regression
  - No guarantee that value of Y is between 0 and 1
  - Cannot handle target variable that follow a Bernoulli distribution with only 2 values

$$Y = \beta_0 + \beta_1 \text{Revenue} + \beta_2 \text{Employees} + \beta_3 \text{VATCompliant}$$

Company	Revenue	Employees	VATCompliant	…	Fraud	Y
ABC	3,000k	400	Y		No	0
BCD	200k	800	N		No	0
CDE	4,2000k	2,200	N		Yes	1
…						
XYZ	34k	50	N		Yes	1

Linear vs Logistic model

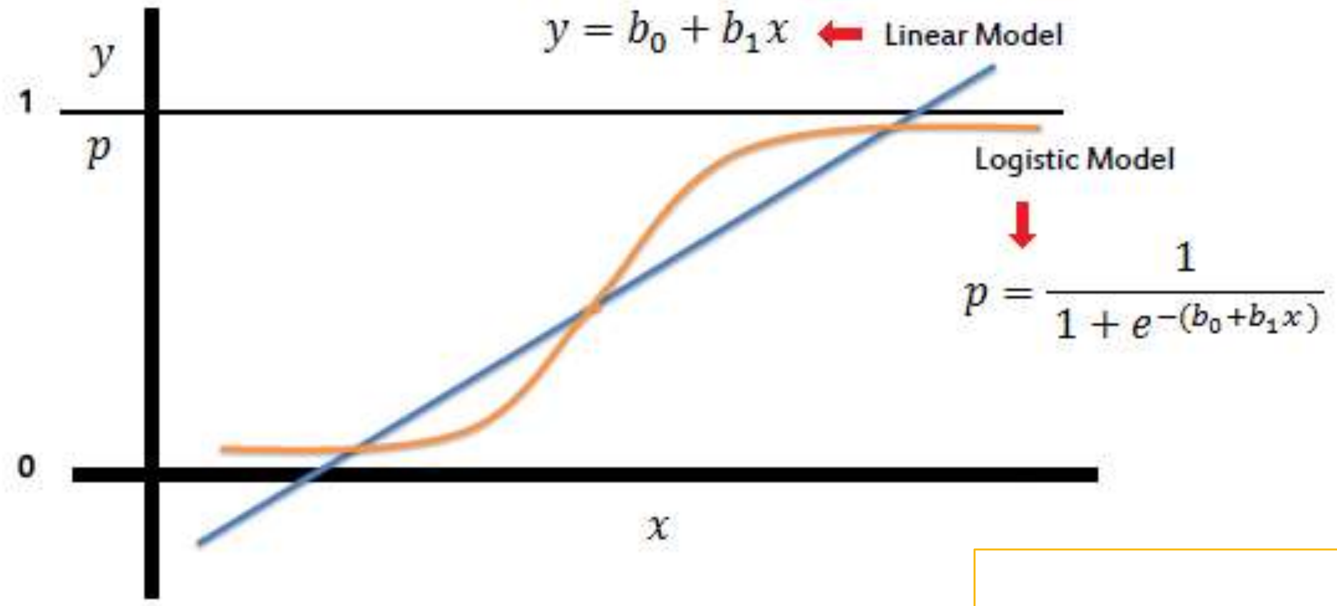


Photo source: https://miro.medium.com/max/1142/1*xTwaKZZsIRek8jzrNWRPzQ.png

Logistic regression can be used for classification problem where the target variable assumes a value between 0 or 1

- ✓ From numerical to binary

$$P(fraud = yes | Revenue,\ Employees,\ VATCompliant) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \text{Revenue} + \beta_2 \text{Employees} + \beta_3 \text{VATcompliant})}}$$

Some basics concepts

- Logistic Regression is a specific type of Generalized Linear Model (GLM) - GLM is a generalization of the concepts and abilities of regular Linear Models

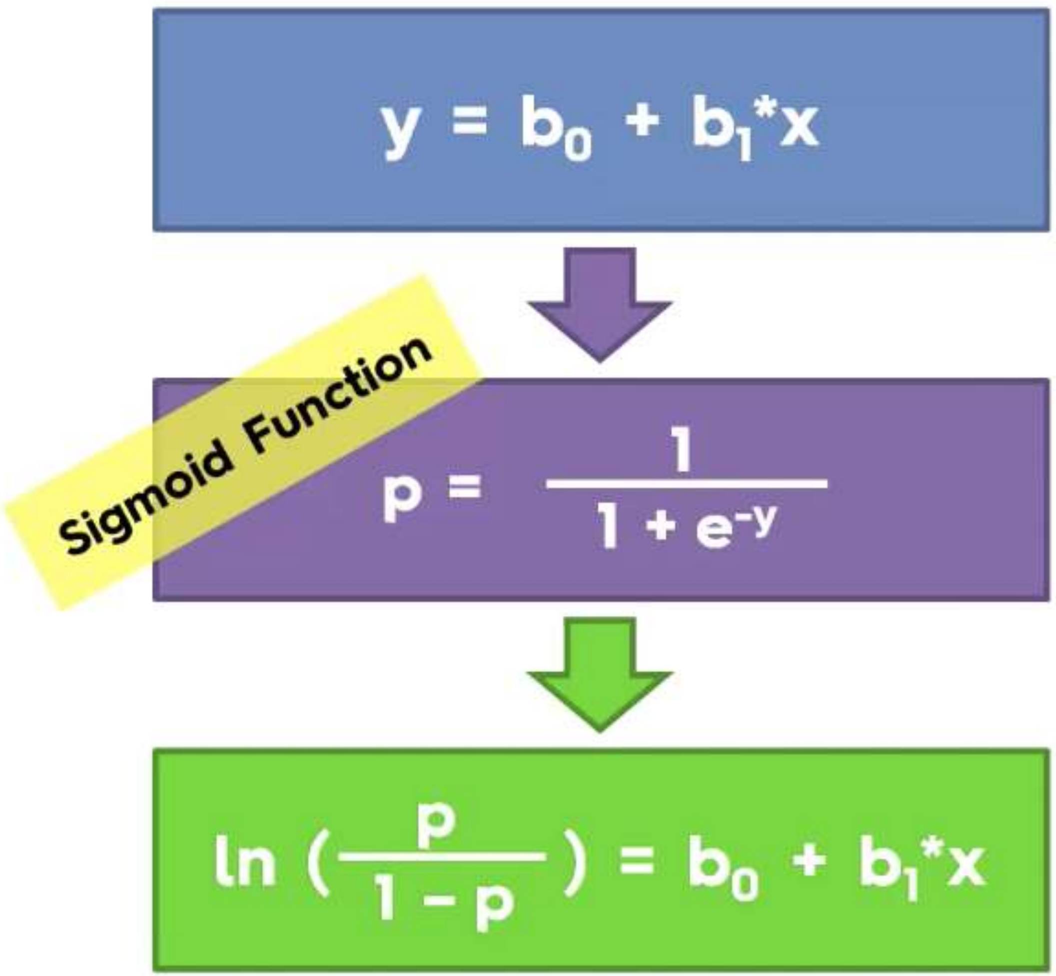
- Assumptions for Logistic Regression
  - No outliers in the data – an outlier can be identified by analysing the independent variables
  - No correlation (multi-collinearity) between the independent variables

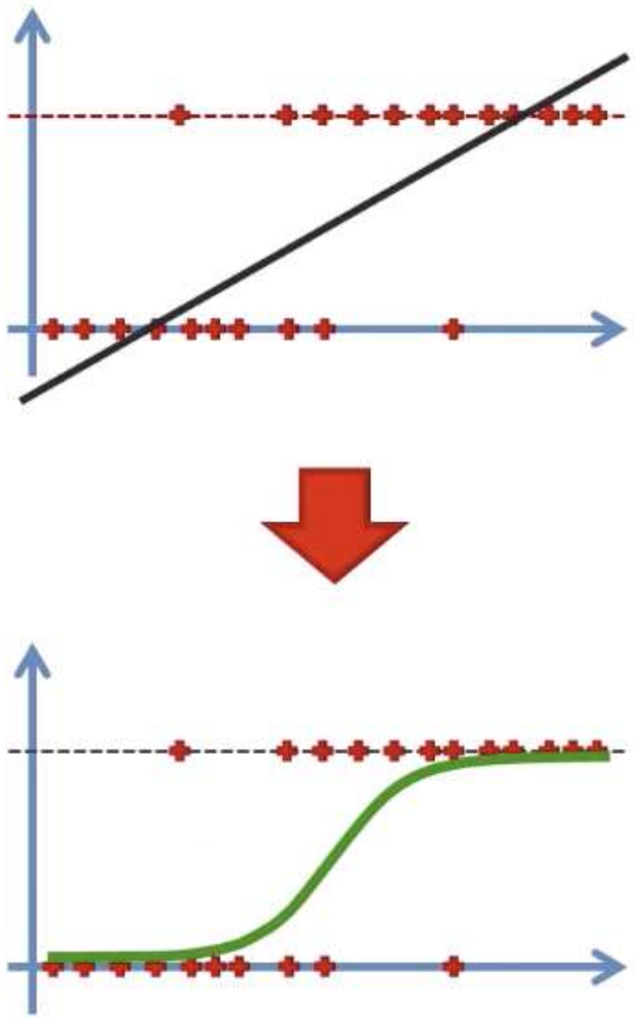
Some basics concepts

- **Probability =** the fraction of times that a fraud happens after many trials, range from 0 to 1
- **Odds** of a fraud = ratio of a fraud happening : a fraud not happening, range from 0 to infinity
- **Log-odds** = logarithm of the odds, or **logit**

- Example : Out of 10 insurance claims, there are 2 fraudulent claims.
- Probability of frauds, P(Y=1) = p = 2/10 = 0.2
  - Thus, Probability of non-frauds, P(Y=0) = 1 – p = 0.8
- Odds of a fraud = 2/8 = 0.25
- p / (1-p) = 0.2/0.8 = 0.25
- Thus, we can express **Odds = p / (1-p)**

Logistic Regression





Picture source: https://miro.medium.com/max/1433/1*rtRMr8B8aoZO44-rfDVnsw.png

Logistic Regression

- Question: How to find the best fit straight line through the data?
- Ans: By optimizing the maximum likelihood estimation (MLE) – choses the parameters in such a way as to maximize the probability of getting the sample at hand

Maximum Likelihood Estimation

For observation i, probability of observing either class:

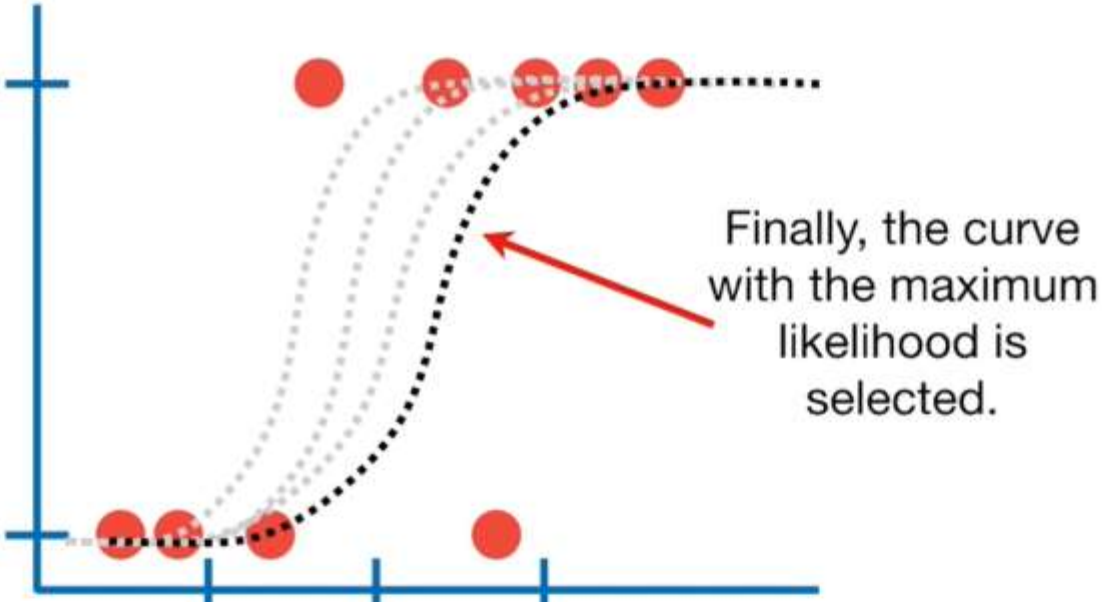
$$P(Y = 1|X_{1i}, \dots, X_{Ni})^{Y_i}(1 - P(Y = 1|X_{1i}, \dots, X_{Ni}))^{1-Y_i},$$

where Y_i represents the target value (either 0 or 1) for observation i

The maximum likelihood function across all n observations:

$$\prod_{i=1}^{n} P(Y = 1|X_{1i}, \dots, X_{Ni})^{Y_i}(1 - P(Y = 1|X_{1i}, \dots, X_{Ni}))^{1-Y_i}$$

Optimize MLE through an iterative process, e.g. Newton method



Picture Source: https://www.youtube.com/watch?v=vN5cNN2-HWE&t=47s

How to interpret Logistic Regression result?

- Logistic Regression
  - linear in log odds (logit)
  - estimates a linear decision boundary between the 2 class (e.g. Fraud vs Legitimate)
- Calculate the odds ratio
  - $\beta_i > 0$ implies $e^{\beta i} > 1$ and the odds and probability increase with X_i
  - $\beta_i < 0$ implies $e^{\beta i} < 1$ and the odds and probability decrease with X_i

where we suppose variable X_i increases with one unit with all other variables being kept constant, then the new logit becomes the old logit with β_i added.

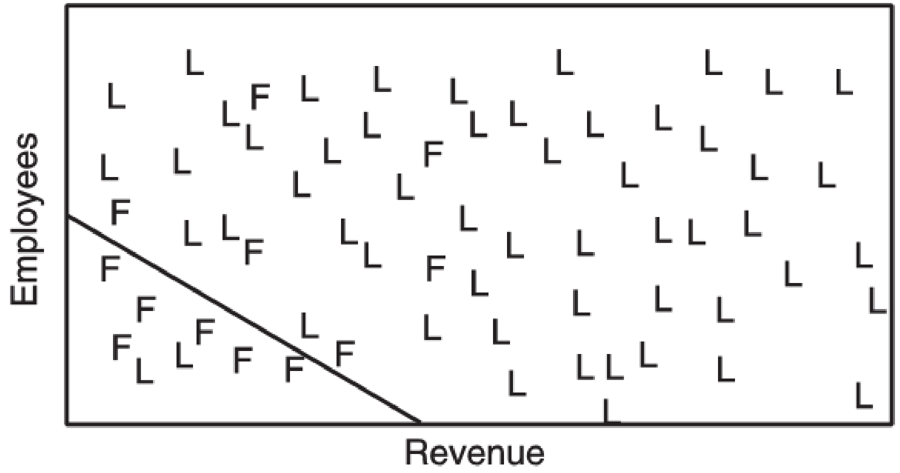


Figure 4.5 Linear Decision Boundary of Logistic Regression

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_N X_N)}}$$

Interpretation is SAME as Linear Regression.

Logistic Regression

Advantages

- Performs exceptionally well for linearly separable data
- Operationally efficient and easy to implement
- A good baseline to measure performance of other more complex fraud detection model

Disadvantages

- Nonlinear problem cannot be solved
- Prone to overfitting
- Difficult to capture complex relationships

Linear vs Logistic Regression

	Linear Regression	**Logistic Regression**
Target variable	Continuous (e.g. claim amount)	Binary (e.g. 0 or 1)
Equation	$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_N X_N$	$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_N X_N)}}$
Purpose	Best fit line (e.g. Ordinary Least Square)	Probability of success or failure of an event (e.g. Maximum Likelihood Estimation)
Output to predict	Continuous value (e.g. $10,000)	Probability (e.g. 0.6, 0.3, 0.9)
Decision	Shows how dependent variable depends on independent variables. Used for prediction.	Helps in decision making. Mainly used for classification purposes based on threshold value.

03
Performance Evaluation

Dr. Vivien CHAN

Evaluating Prediction Models

- Mainly involves 2 steps
- 1/ Split the sample data
  - Training data
  - Validation data
  - Testing data
- 2/ Model evaluation
  - Classification model
  - Regression model

Sample Data Set

- Split the sample data set into 2-3 datasets

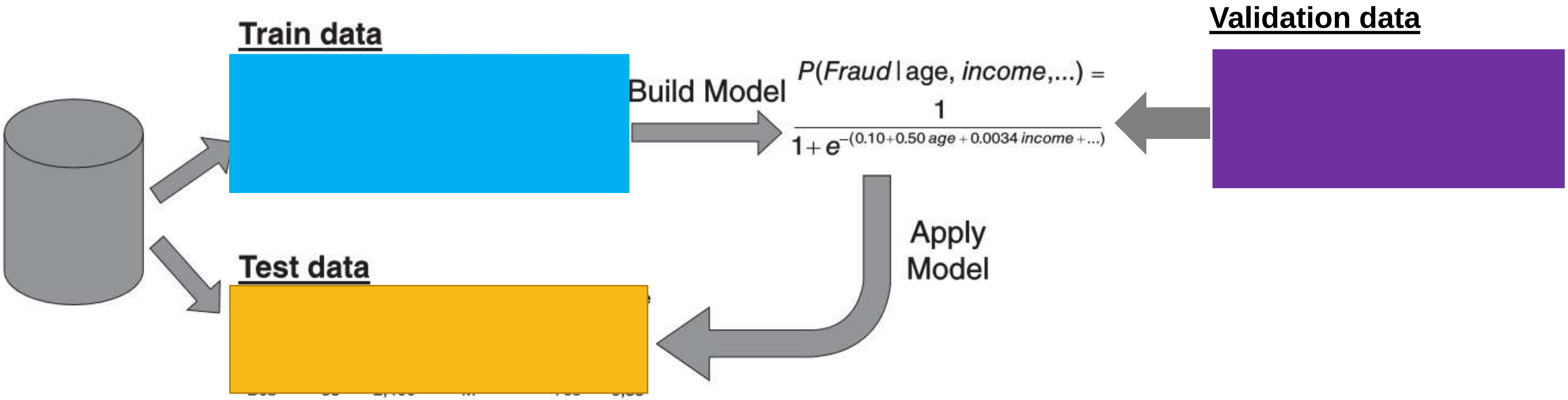
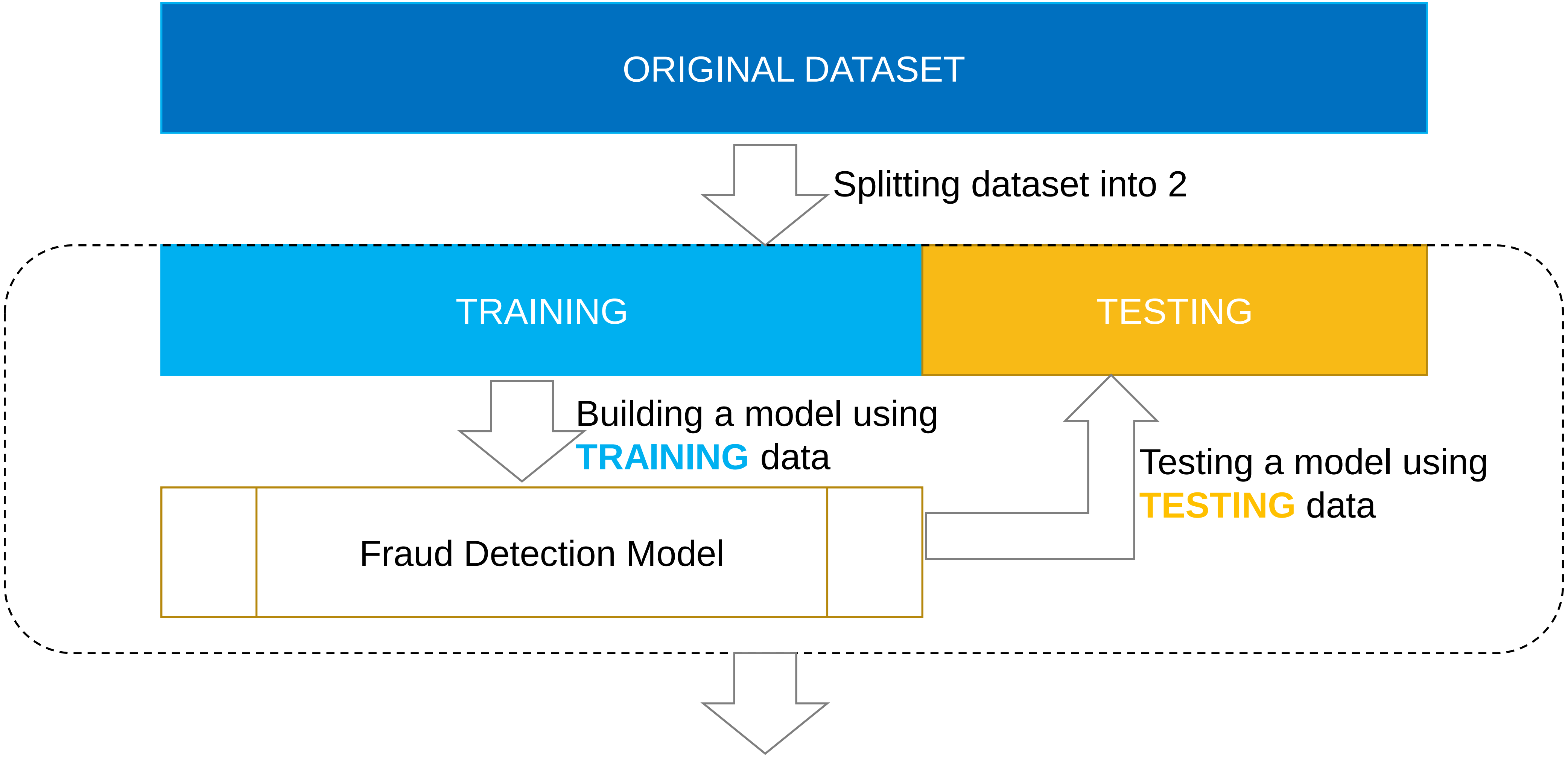


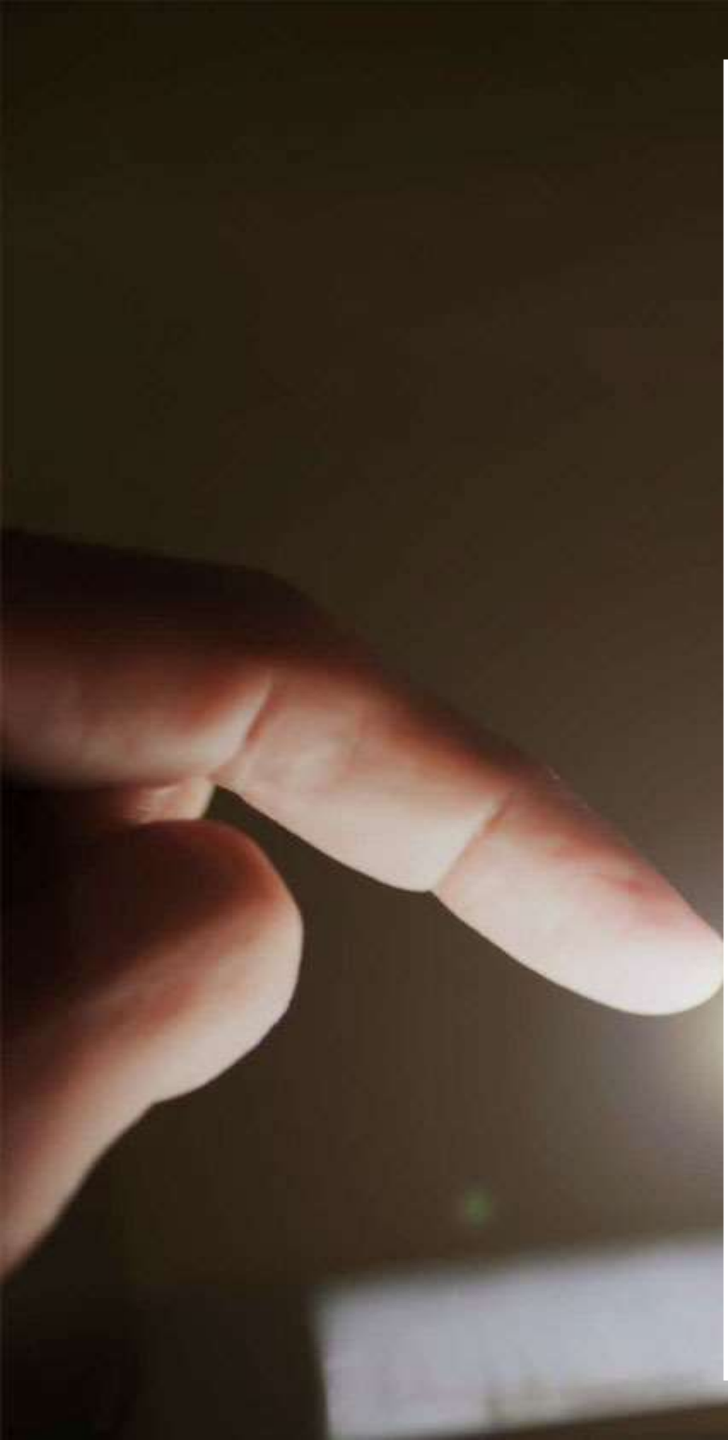
Figure 4.34 Training Versus Test Sample Set Up for Performance Estimation

Source: Fraud Analytics using Descriptive, Predictive, and Social Network Techniques, 1st ed, John Wiley & Sons Inc.

Scenario 1

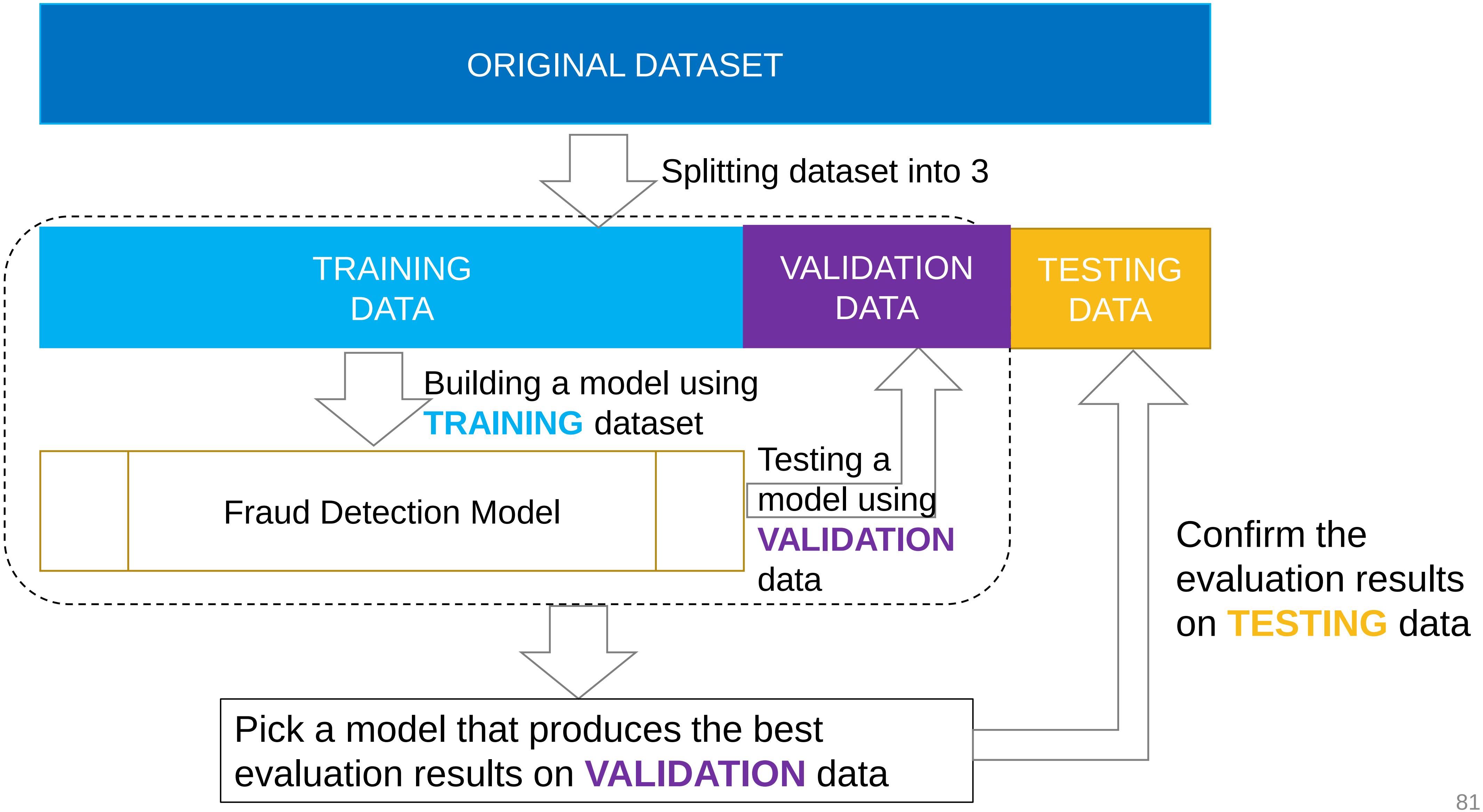


Pick a model that produces the best
evaluation results on TESTING data



If we can train a model using the training data and evaluate it using the testing data. So, Why do we need a validation data set?

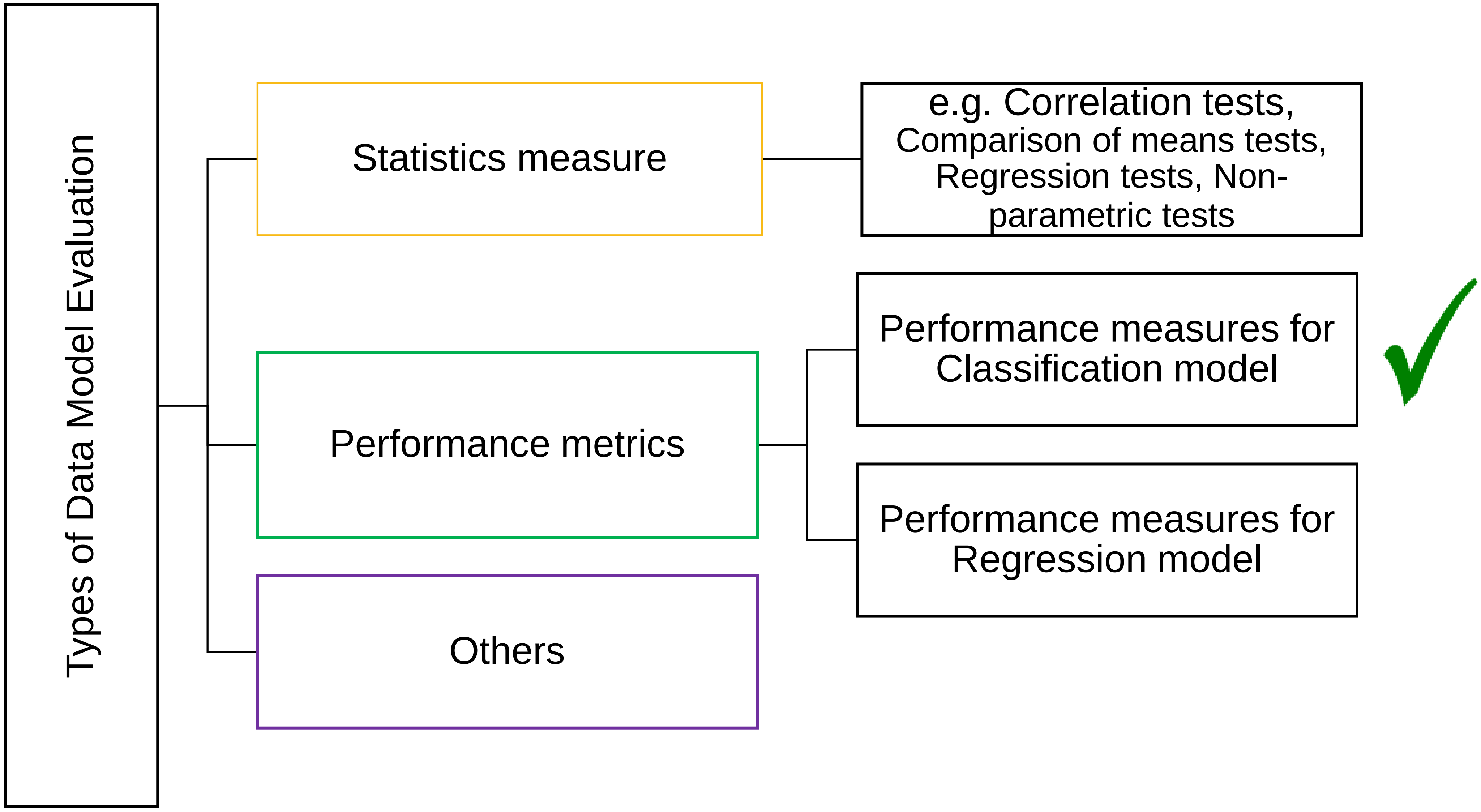
Scenario 2


ORIGINAL DATASET
Splitting dataset into 3
TRAINING DATA
VALIDATION DATA
TESTING DATA
Building a model using TRAINING dataset
Fraud Detection Model
Testing a model using VALIDATION data
Confirm the evaluation results on TESTING data
Pick a model that produces the best evaluation results on VALIDATION data

Training vs Validation vs Testing

Training data	**Purpose**: to build the model
	A dataset of observations used during the learning process
	The goal is to produce a trained (fitted) model that generalizes well to new, unknown data
	Training data should not be used for validation or testing
Validation data	**Purpose**: to be used during model development (e.g. making stopping decision in decision tree)
	A dataset of observations used to tune the hyperparameters of a prediction model (e.g. decision of when to stop growing a decision tree)
	Training stopped with the minimum error on the validation set
Testing data	**Purpose** : to test the performance of the model
	A dataset of observations independent of the training and validation datasets
	Used only to assess the performance of a trained prediction model
	Overfitting – a better fitting of the training dataset than the testing dataset

Types of Data Model Evaluation



Confusion Matrix

- The Confusion Matrix – an example

	Fraud	Fraud Score
John	Yes	0.72
Sophie	No	0.56
David	Yes	0.44
Emma	No	0.18
Bob	No	0.36

Cut-Off = 0.50

	Fraud	Fraud Score	Predicted
John	Yes	0.72	Yes
Sophie	No	0.56	Yes
David	Yes	0.44	No
Emma	No	0.18	No
Bob	No	0.36	No

Figure 4.37 Calculating Predictions Using a Cut-Off

Table 4.5 Confusion Matrix

		Actual Status	
		Positive (Fraud)	Negative (No Fraud)
Predicted status	Positive (Fraud)	True Positive (John)	False Positive (Sophie)
	Negative (No Fraud)	False Negative (David)	True Negative (Emma, Bob)

Source: Fraud Analytics using Descriptive, Predictive, and Social Network Techniques, 1st ed, John Wiley & Sons Inc.

Accuracy

		Actual	
		Fraud	No fraud
Prediction	Fraud	True Positive (TP)	False Positive (FP)
	No fraud	False Negative (FN)	True Negative (TN)

$$\text{Classification accuracy} = (TP + TN)/(TP + FP + FN + TN)$$

Accuracy: Percentage of total items classified correctly

$$\text{Classification error} = (FP + FN)/(TP + FP + FN + TN)$$

Error: Percentage of total items classified incorrectly

Example

		Actual	
		Fraud	No fraud
Prediction	Fraud	True Positive (TP) = 0	False Positive (FP) = 0
	No fraud	False Negative (FN) = 10	True Negative (TN) = 90

Aaccuracy = (0+90) / 100 = 90%

Even with very high accuracy, this model is useless in detecting fraud cases.

Recall

		Actual	
		Fraud	No fraud
Prediction	Fraud	True Positive (TP)	False Positive (FP)
	No fraud	False Negative (FN)	True Negative (TN)

$$\text{Sensitivity} = \text{Recall} = \text{Hit rate} = TP/(TP + FN)$$

Recall: measures how many fraudsters are correctly classified as fraudsters

This is the most important performance measure for fraud detection models, i.e. favour TP > FN

Precision

		Actual	
		Fraud	No fraud
Prediction	Fraud	True Positive (TP)	False Positive (FP)
	No fraud	False Negative (FN)	True Negative (TN)

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$

Precision: measures how many predicted fraudsters are actually fraudsters

This is useful measure if the objective is not to leave out important information, e.g. spam mail detection. That means you would like to have TP > FP

F1 score

		Actual	
		Fraud	No fraud
Prediction	Fraud	True Positive (TP)	False Positive (FP)
	No fraud	False Negative (FN)	True Negative (TN)

$$\text{F-measure} = 2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$$

F1 score: the weighted average of Precision and Recall

This takes into account FP and FN, thus more informative than accuracy.

Example

		Actual	
		Fraud	No fraud
Prediction	Fraud	True Positive (TP) = 2	False Positive (FP) = 3
	No fraud	False Negative (FN) = 8	True Negative (TN) = 87

Accuracy = (2+87) / 100 = 89%
Recall = (2)/(2+8)=20%
Precision = (2)/(2+3)=40%
F1 score = 2x20%x40%/(20%+40%) = 27%

How to interpret these 4 measures?

For fraud detection models, Recall is most useful performance measure.

TNR, FPR

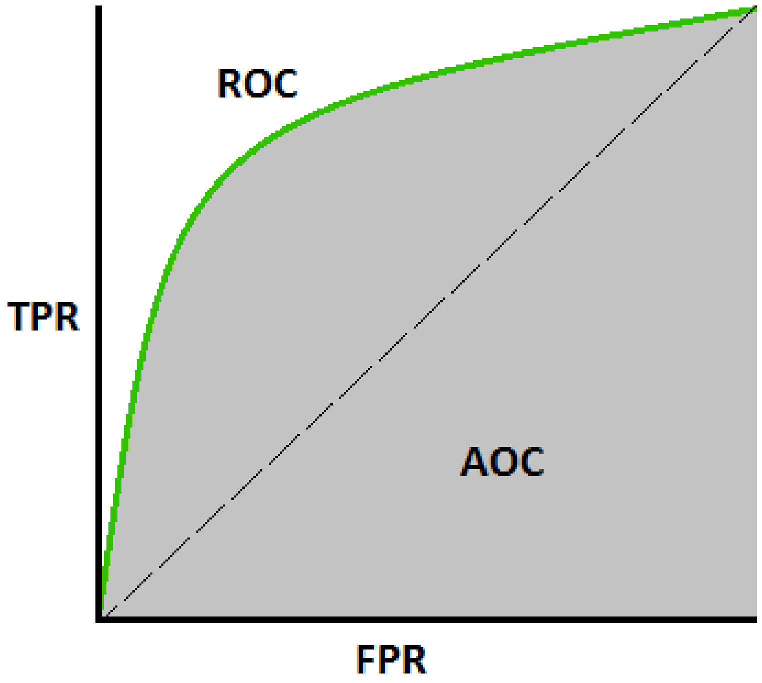
		Actual	
		Fraud	No fraud
Prediction	Fraud	True Positive (TP)	False Positive (FP)
	No fraud	False Negative (FN)	True Negative (TN)

True Negative Rate (TNR) = Specificity = TN / (FP+TN)

False Positive Rate (FPR) = 1-Specificity = FP / (FP+TN)

ROC-AUC

- Some basic terms:
  - ROC = Receiver Operating Characteristics
  - AUC = Area under the ROC curve
  - True Positive Rate (TPR) = Sensitivity = Recall = Hit rate = TP/(TP+FN)
  - True Negative Rate (TNR) = Specificity = TN/(FP+TN)
  - False Positive Rate (FPR) = 1-Specificity = FP/(FP+TN)
- What is ROC curve?
  - It is a curve of probabilities,
  - with TPR as y-axis and FPR as x-axis



ROC-AUC : Example

- If we use different “Cut-off” as the threshold, we’ll have different prediction for “Fraud” and “No Fraud” cases

	Fraud	Fraud Score
John	Yes	0.72
Sophie	No	0.56
David	Yes	0.44
Emma	No	0.18
Bob	No	0.36

Cut-Off = 0.50 →

	Fraud	Fraud Score	Predicted
John	Yes	0.72	Yes
Sophie	No	0.56	Yes
David	Yes	0.44	No
Emma	No	0.18	No
Bob	No	0.36	No

Figure 4.37 Calculating Predictions Using a Cut-Off

For example,
If use “Cut-off = 0.40”,
John, Sophie and David will be classified as “Fraud” case.

If use “Cut-off = 0.60”,
only John will be classified as “Fraud” case.

Source: Fraud Analytics using Descriptive, Predictive, and Social Network Techniques, 1st ed, John Wiley & Sons Inc.

Example

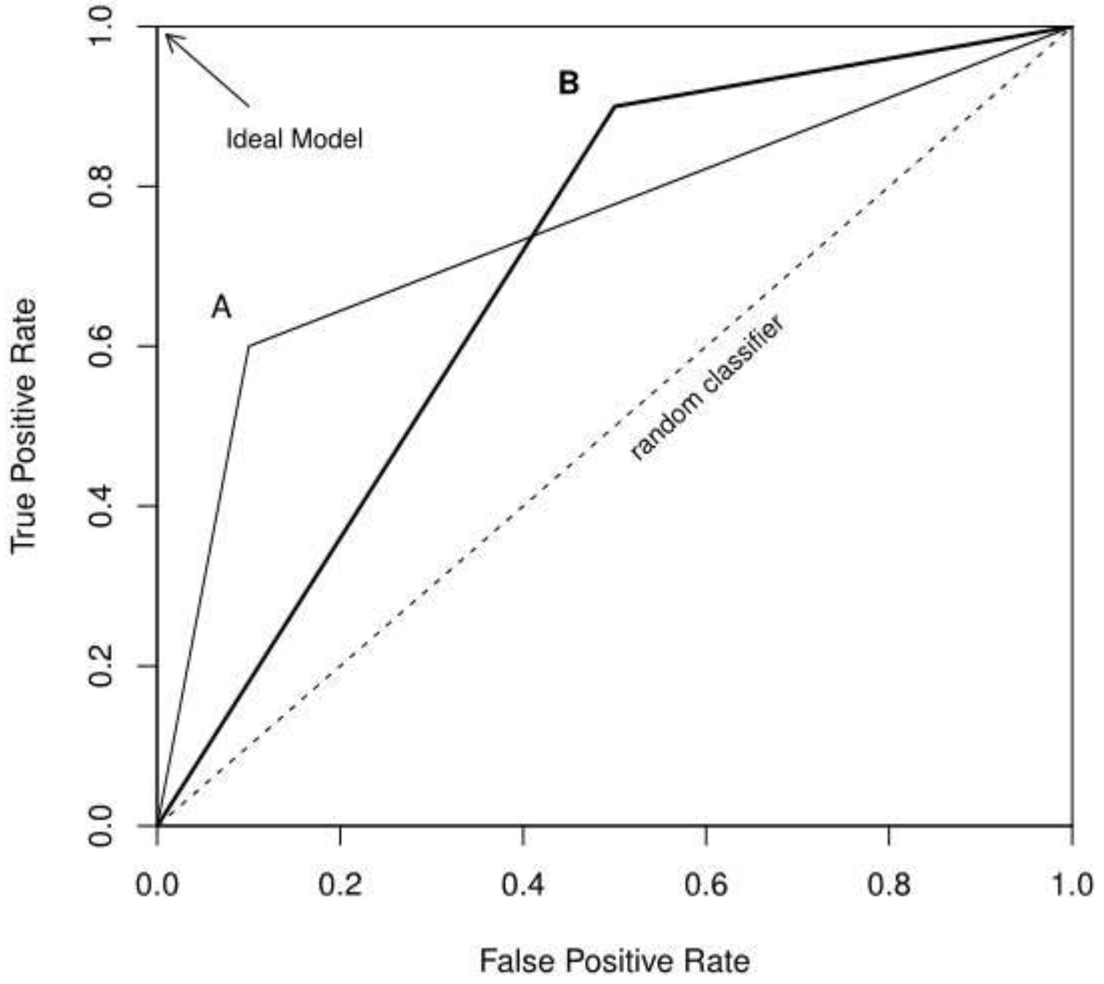
Table 4.6 Table for ROC Analysis

Cut-off	Sensitivity	Specificity	1-Specificity
0	1	0	1
0.01			
0.02			
….			
0.99			
1	0	1	0

- For example, if “cut-off = 0.5”
  - Sensitivity = 50%
  - Specificity = 67%
  - 1-Specificity = 33%

Source: Fraud Analytics using Descriptive, Predictive, and Social Network Techniques, 1st ed, John Wiley & Sons Inc.

Example



- AUC is always between 0 and 1
- AUC = 1 = ideal situation where all fraud and no fraud cases are correctly predicted
- AUC = 0.5 (i.e. diagonal curve) = random guesses, i.e. no discrimination power between fraud and no fraud cases
- Any curves under the diagonal curve = no use


CASE
STUDY
VectorStock

Case Background

- From a study by Sabău et al (2021)
- The sample was selected from the Bucharest Stock Exchange (i.e. a Romania Stock Exchange) and consists of 66 companies traded on the main market, for the years 2015–2019.
- Objective of the study:
  - to identify which of the eight-variables from the Beneish model that influences the most or least the outcome of the final score
  - What are the financial indicators that most strongly discriminate the two states: fraudulent financial statements and non-fraudulent financial statements?

Case Background

- What is Beneish model?
  - Created by Professor M. Daniel Beneish of the Kelley School of Business at Indiana University
  - A mathematical model that uses financial ratios and eight variables to identify whether a company has manipulated its earnings. It is used as a tool to uncover financial fraud.
  - Eight variables:
    - Days Sales in Receivables Index (DSRI) ,Gross Margin Index (GMI), Asset Quality Index (AQI), Sales Growth Index (SGI), Depreciation Index (DEPI), Sales General and Administrative Expenses Index (SGAI), Leverage Index (LVGI), Total Accruals to Total Assets (TATA)

The M-Beneish equation is as follows:

$$M = -4.84 + 0.92 \times DSRI + 0.528 \times GMI + 0.404 \times AQI + 0.892 \times SGI + 0.115 \times DEPI - 0.172 \times SGAI + 4.679 \times TATA - 0.327 \times LVGI$$

NOTE: Details of Beneish Model will not be covered in this course

Case Background

- Companies are categorized in to FRAUD and NON-FRAUD based on Beneish score using a threshold value assigned by the author
- Logistic Regression is applied afterwards

Result interpretation

- Logistic Regression result

Table 3. Univariate binary logistic regression.

Variable	Univariate Logistic Regression		ROC Curve	
	Exp(B)	*p*-Value	AUROC	*p*-Value
DSRI	1.815	0.329	0.536	0.634
GMI	8.316	0.007	0.854	0.000
AQI	6.183	0.047	0.686	0.014
SGI	1.459	0.683	0.493	0.924
DEPI	1.687	0.057	0.670	0.025
SGAI	1.056	0.545	0.465	0.644
LVGI	5.536	0.222	0.580	0.295
TATA	89.801	0.053	0.752	0.001

How to interpret this logistic regression result?

References

- Bart Baesens, Veronique Van Vlasselaer, Wouter Verbeke (2015). Fraud Analytics using Descriptive, Predictive, and Social Network Techniques, 1st ed, John Wiley & Sons Inc.
- Leonard W. Vona (2017). Fraud Data Analytics Methodology: The Fraud Scenario Approach to Uncovering Fraud in Core Business Systems, John Wiley & Sons, Inc.
- Sabău (Popa), Andrada-Ioana, Codruța Mare, and Ioana Lavinia Safta (2021). A Statistical Model of Fraud Risk in Financial Statements. Case for Romania Companies, Risks 9, no. 6: 116. https://doi.org/10.3390/risks9060116
- Spann, Delena D.. (2013). Fraud Analytics : Strategies and Methods for Detection and Prevention, John Wiley & Sons, Incorporated, 2013. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/hkuhk/detail.action?docID=1752695

QUESTIONS?