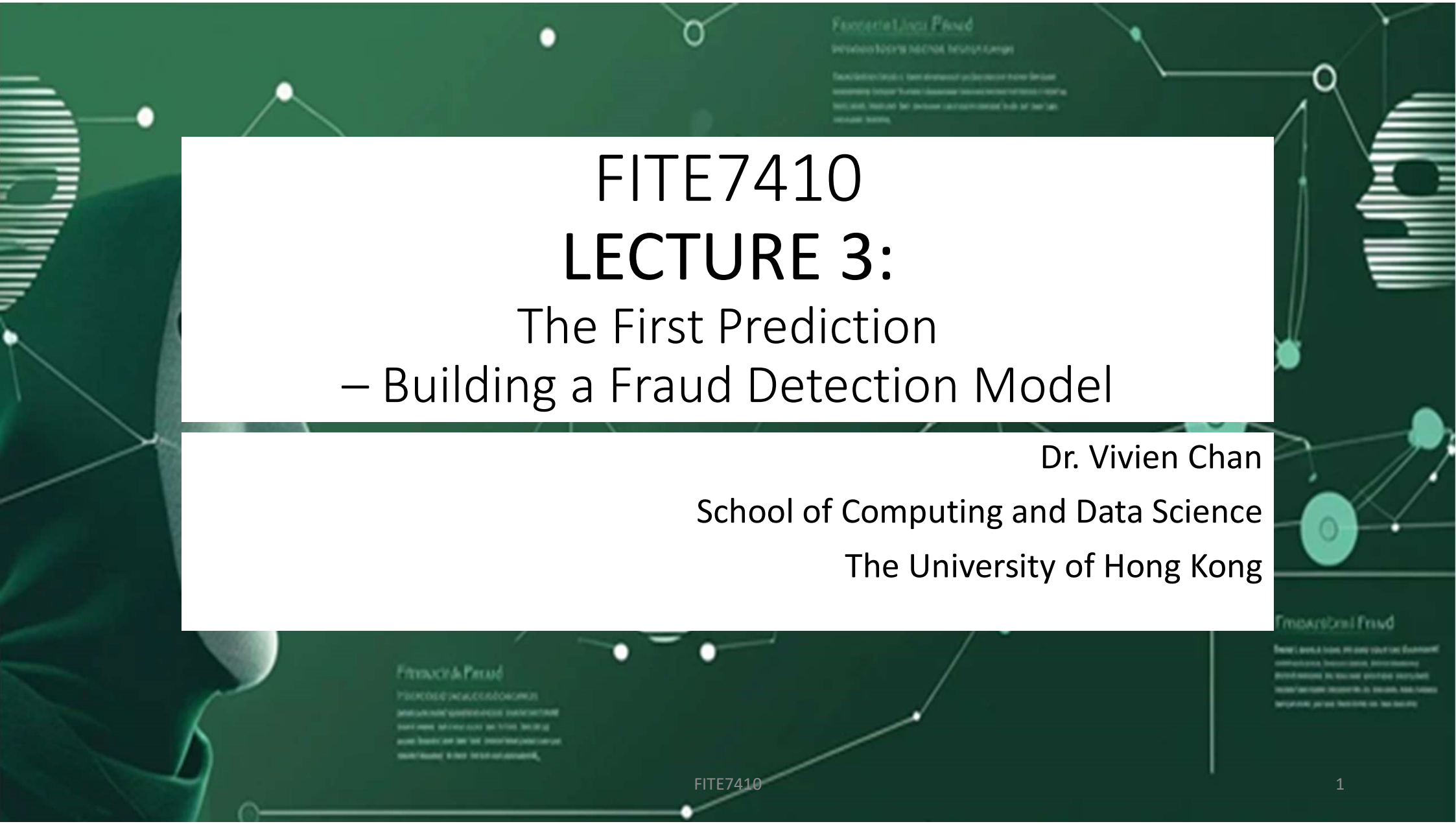


# FITE7410
# LECTURE 3:
## The First Prediction
## – Building a Fraud Detection Model

Dr. Vivien Chan

School of Computing and Data Science

The University of Hong Kong

# Agenda

Lecture goals:

1) Understand the Red Flags related to Financial Statement Frauds.
2) Learn the first prediction model – Linear and Logistic Regression.
3) Understand the use of key performance metrics for fraud detection models.


A fintech company needs an automated system to flag suspicious transactions. How do we build a model that is not just accurate, but *effective?*

# The New Problem: Financial Statement Fraud

# 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/

# 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.

# Example of data samples

- Problem with limited number of fraud cases

| | |
|---|---|
| **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.

# Some examples of 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.

# Examples of financial attributes – Beneish model

- 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:

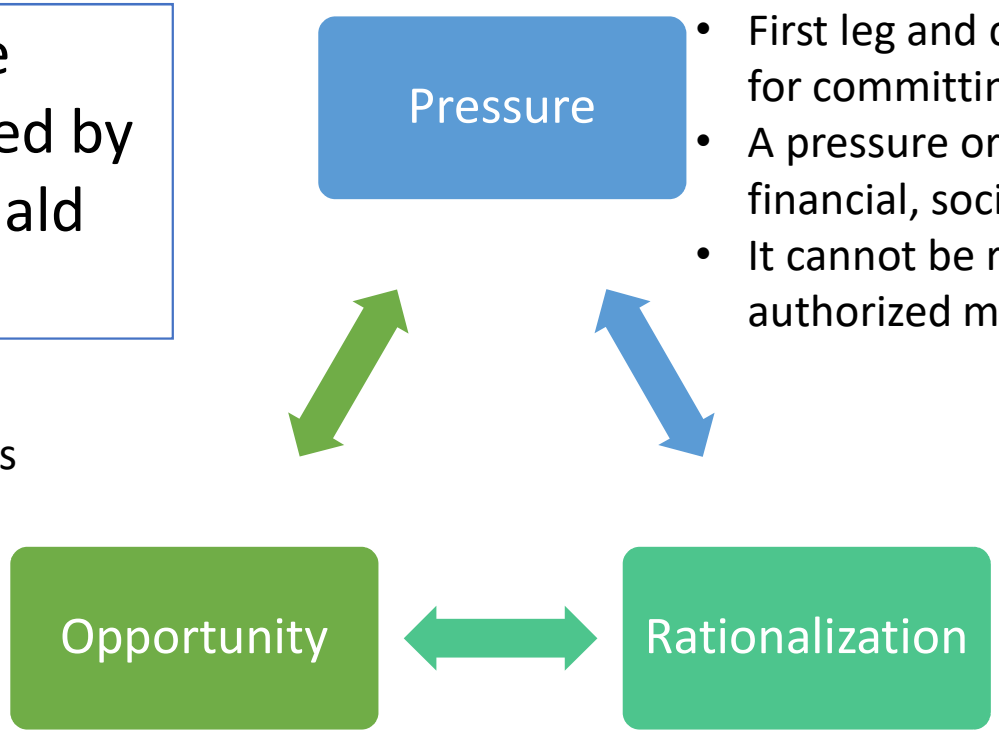
$$M = -4.84 + 0.92 \times DSRI + 0.528 \times GMI + 0.404 \times AQI + 0.892 \times SGI + 0.115 \times DEPI$$

$$- 0.172 \times SGAI + 4.679 \times TATA - 0.327 \times LVGI$$

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

# Fraud Triangle

The Fraud Triangle Theory is developed by criminologist, Donald Cressey

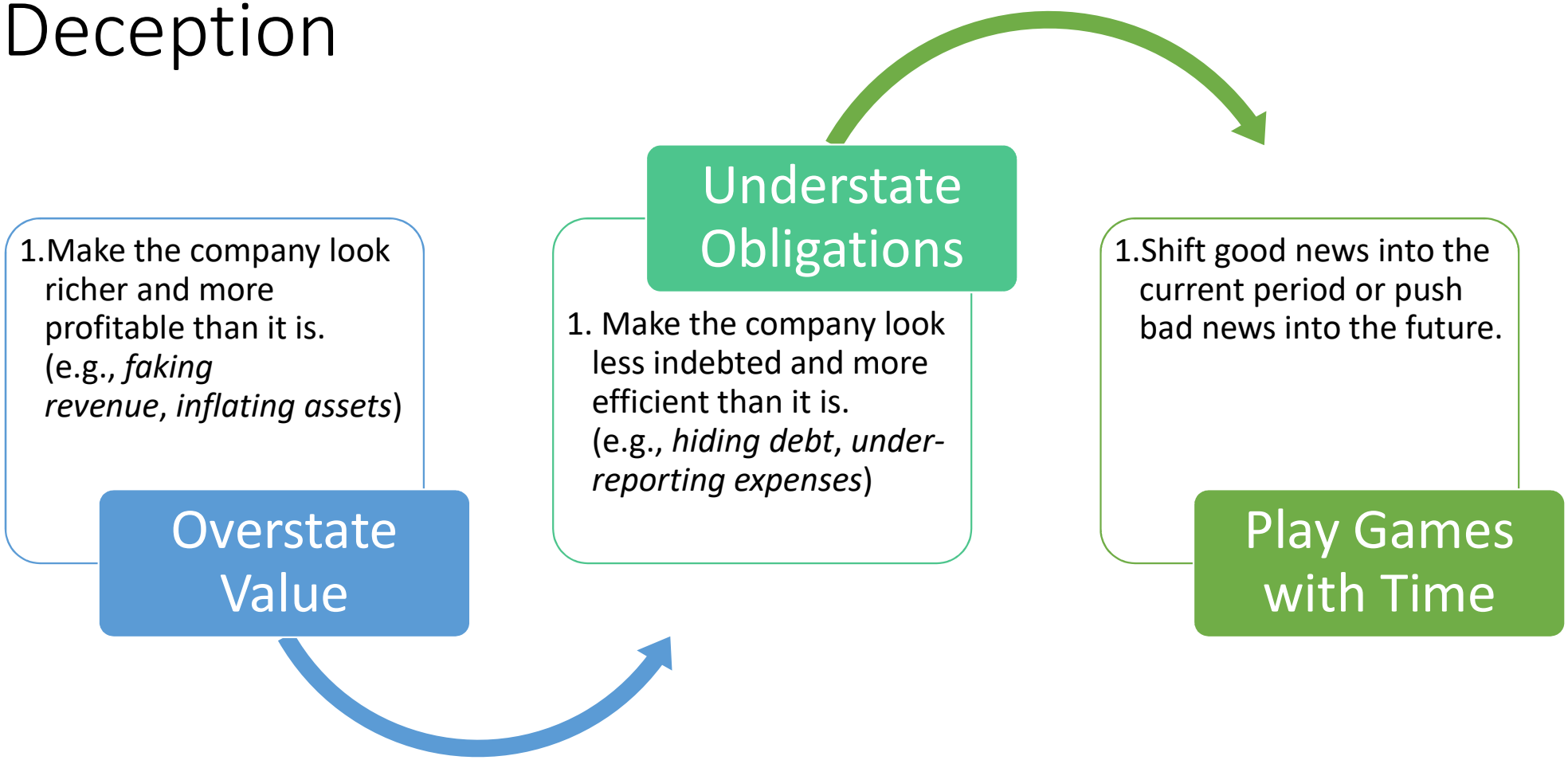



- First leg and concerns the main motivation for committing fraud
- A pressure or a problem is experienced of financial, social, or any other nature
- It cannot be resolved or relieved in an authorized manner

- Second leg and concerns the precondition for committing fraud
- Opportunity exists to resolve or relieve the experienced pressure in an unauthorized but concealed or hidden manner

- Psychological mechanism that explains why fraudsters do not refrain from committing fraud and think of their conduct as acceptable

# The Three Core Strategies of Financial Deception



# Example techniques

- 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

# Hypothesis #1: Are They Faking Revenue?


A company records fake sales to customers who don't exist or recognizes revenue before it's actually earned.

The Analyst's Question

If they were booking sales but no real cash was coming in, what part of the financial statements would look strange?

The Data Clue (The First Red Flag)

Accounts Receivable: This account, which tracks money owed to the company by customers, would grow much faster than actual sales.

The Testable Hypothesis

A sharp, unexplained increase in the Days' Sales in Receivables Index (DSRI) is a strong indicator of potential revenue manipulation.

# Hypothesis #2: Is the Business Weaker Than It Looks?


The company's core profitability is declining, so management uses accounting tricks to hide this deterioration. They might improperly capitalize costs (turning expenses into assets) or hide debt.

The Analyst's Question

If the underlying quality of the business is getting worse, where would we see the strain?

The Data Clue (The First Red Flag)

- Gross Margin Index (GMI): Measures if the company's core profitability is shrinking. A GMI greater than 1 is a powerful indicator of motive to commit fraud.
- Asset Quality Index (AQI): Measures the proportion of "soft" assets versus hard assets. A spike here suggests the company might be turning expenses into fictitious assets to inflate profits.

The Testable Hypothesis

- A GMI greater than 1 is a powerful indicator of motive to commit fraud.
- A spike in AQI suggests the company might be turning expenses into fictitious assets to inflate profits.

# An example: From Hypotheses to an Investigative Checklist

| **Investigative Question** | **Fraud Scheme** | **Key Forensic Metric** | **What a Red Flag Looks Like** |
|---|---|---|---|
| *Are they faking revenue?* | Inflating or accelerating revenue | Days' Sales in Receivables Index (DSRI) | A sharp increase (Index > 1) |
| *Is the business quality poor?* | Hiding deteriorating margins | Gross Margin Index (GMI) | A GMI > 1 suggests shrinking margins |
| | Improperly capitalizing costs | Asset Quality Index (AQI) | A sharp increase (Index > 1) |
| *Is growth unsustainable?* | Desperate push for sales | Sales Growth Index (SGI) | A high index is a contextual risk |
| | Manipulating non-cash expenses | Depreciation Index (DEPI) | A DEPI > 1 suggests depreciation has slowed |

# The First Fraud Detection Model

Linear and Logistic Regression

# 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)

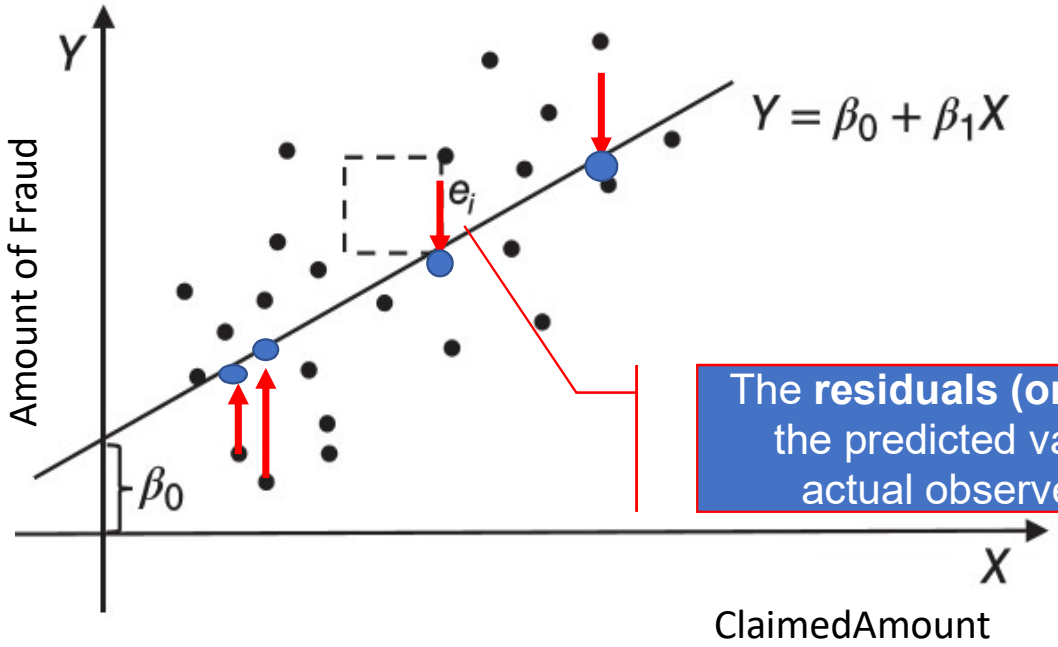
$X_{1, \ldots} X_N$ = explanatory variables (or independent variables)

$\beta$ = 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)

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




MSE = mean squared error

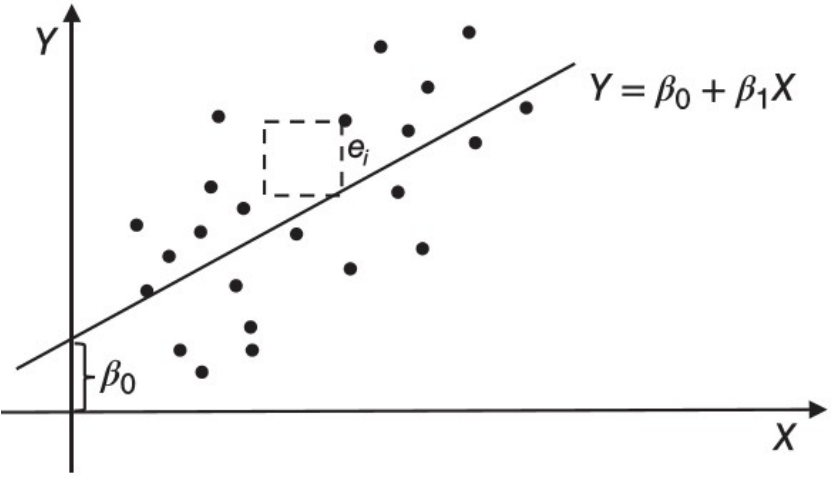
$n$ = number of data points

$Y_i$ = 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

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

Example:

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

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

$\beta_{1\ldots}\beta_N$ = 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.

$\beta_0$ = Intercept coefficient
i.e. expected mean value of Y when all X=0. However, if X never = 0, Y will have no meaning.

# 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)

# MAE, MSE, RMSE

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

- MAE
  - where $\mathbf{y_i}$ is the actual expected output and $\hat{\mathbf{y}}_\mathbf{i}$ 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.

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

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

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

- 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

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

- 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.

# Additional Performance Measures for Logistic Regression - Visualization

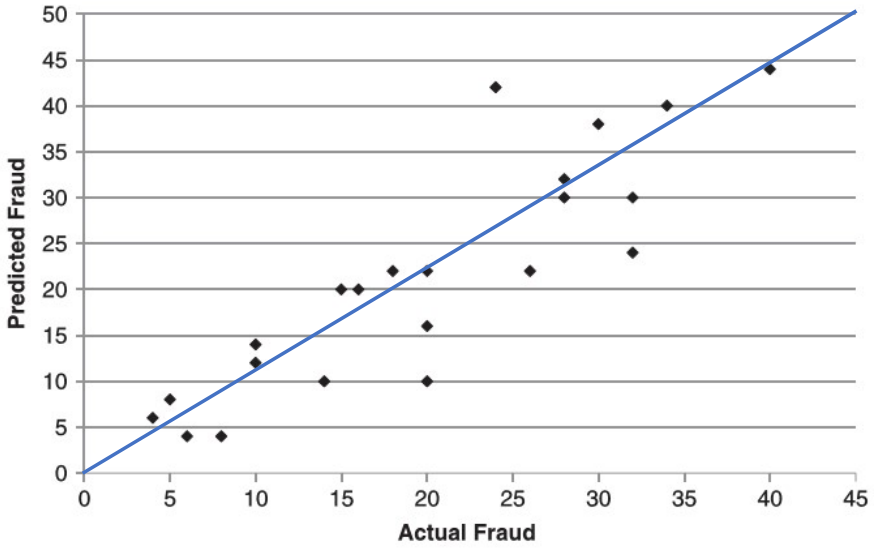



**Figure 4.44** Scatter Plot: Predicted Fraud Versus Actual Fraud

- 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 $i$, $\bar{\hat{y}}$ 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.

# Additional Performance Measures for Logistic Regression

- Akaike Information Criterion (AIC): A measure of the model's fit while penalizing model complexity. A lower AIC indicates a better fit with fewer parameters.

$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

k = number of parameters in the model
L^ = maximized value of the likelihood function of the model

- Bayesian Information Criterion (BIC): Similar to AIC but with a stronger penalty for model complexity. A lower BIC indicates a better fit with fewer parameters.

$$\mathrm{BIC} = \ln(n)k - 2\ln(\hat{L}).$$

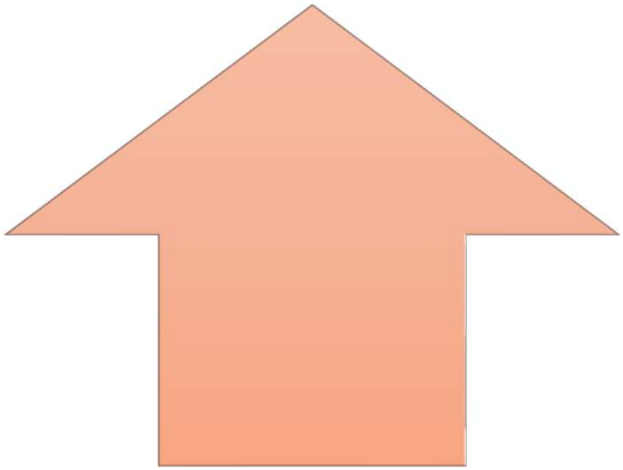
n = number of data points
k = number of parameters in the model
L^ = maximized value of the likelihood function of the model

# Interpreting R-squared, AIC, and BIC for Model Fit

- Fill in the table below

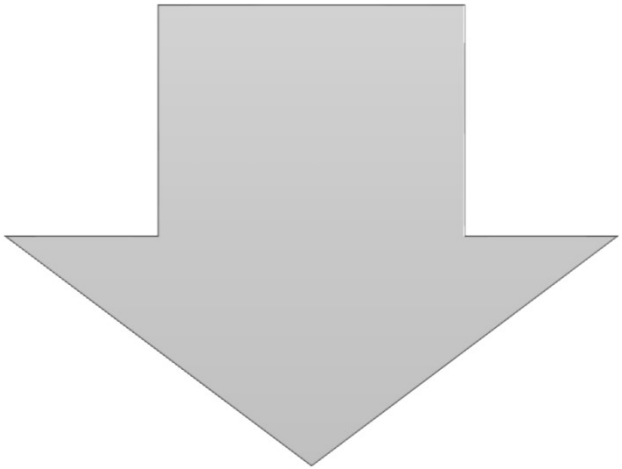
| *Metric* | *Higher or Lower for Good Fit?* |
|---|---|
| R-squared | |
| AIC | |
| BIC | |

# Linear Regression



## Advantages

- Interpretability
- Simplicity
- Wide Applicability
- Computational Efficiency
- Baseline Performance



## Disadvantages

- Limited Expressiveness - Target and exploratory variables must be of linear relation
- Prone to noise and overfitting
- Sensitive to outliers
- Assumptions: linearity, normality, homoscedasticity

# Logistic Regression

- Limitations of 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

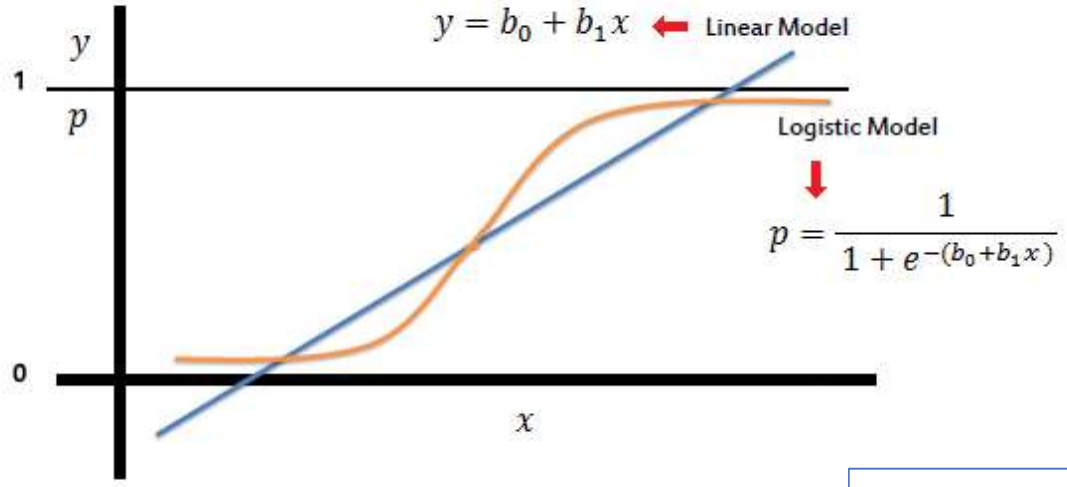



Photo source: https://miro.medium.com/max/1142/1*xTwaKZZsIRek8jzrNWRPzQ.png

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

**Probability (p)**

**=** the fraction of times that a fraud happens after many trials, range from 0 to 1

**Odds** of a fraud (**p / (1-p)**

= ratio of a fraud happening : a fraud not happening, range from 0 to infinity

**Log-odds** = logarithm of the odds, or **logit (ln (p / (1-p)) )**

- 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, i.e. p / (1-p) = 0.2/0.8 = 0.25
- Thus, we can express **Odds = p / (1-p)**

# 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 $\beta_i$ added.

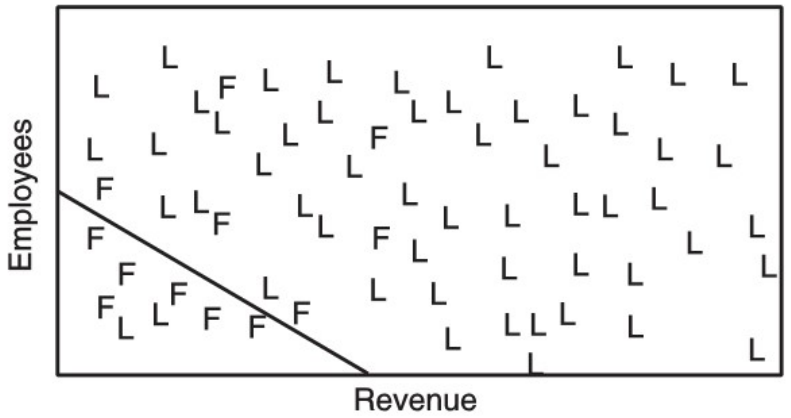
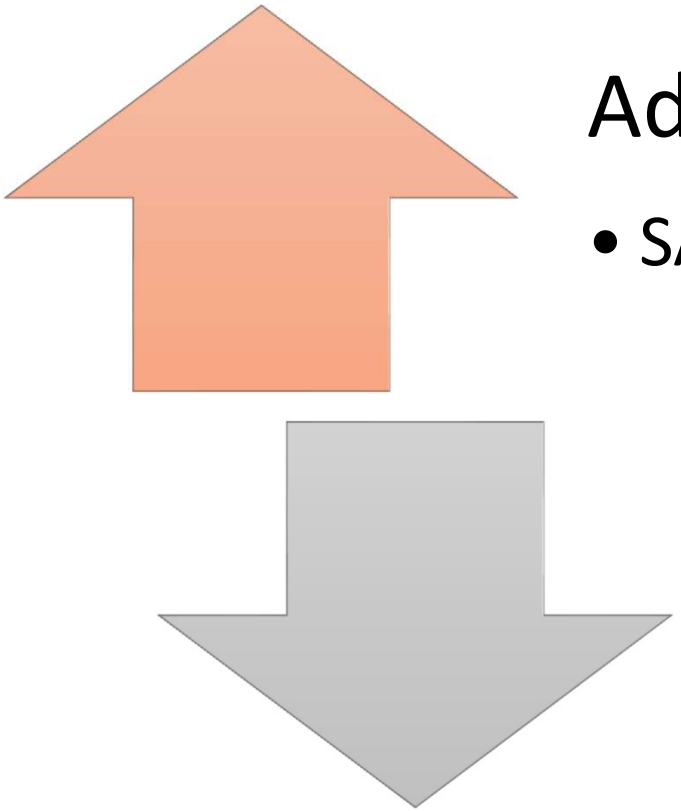



**Figure 4.5** Linear Decision Boundary of Logistic Regression

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

Interpretation is SAME as Linear Regression.

# Logistic Regression



## Advantages

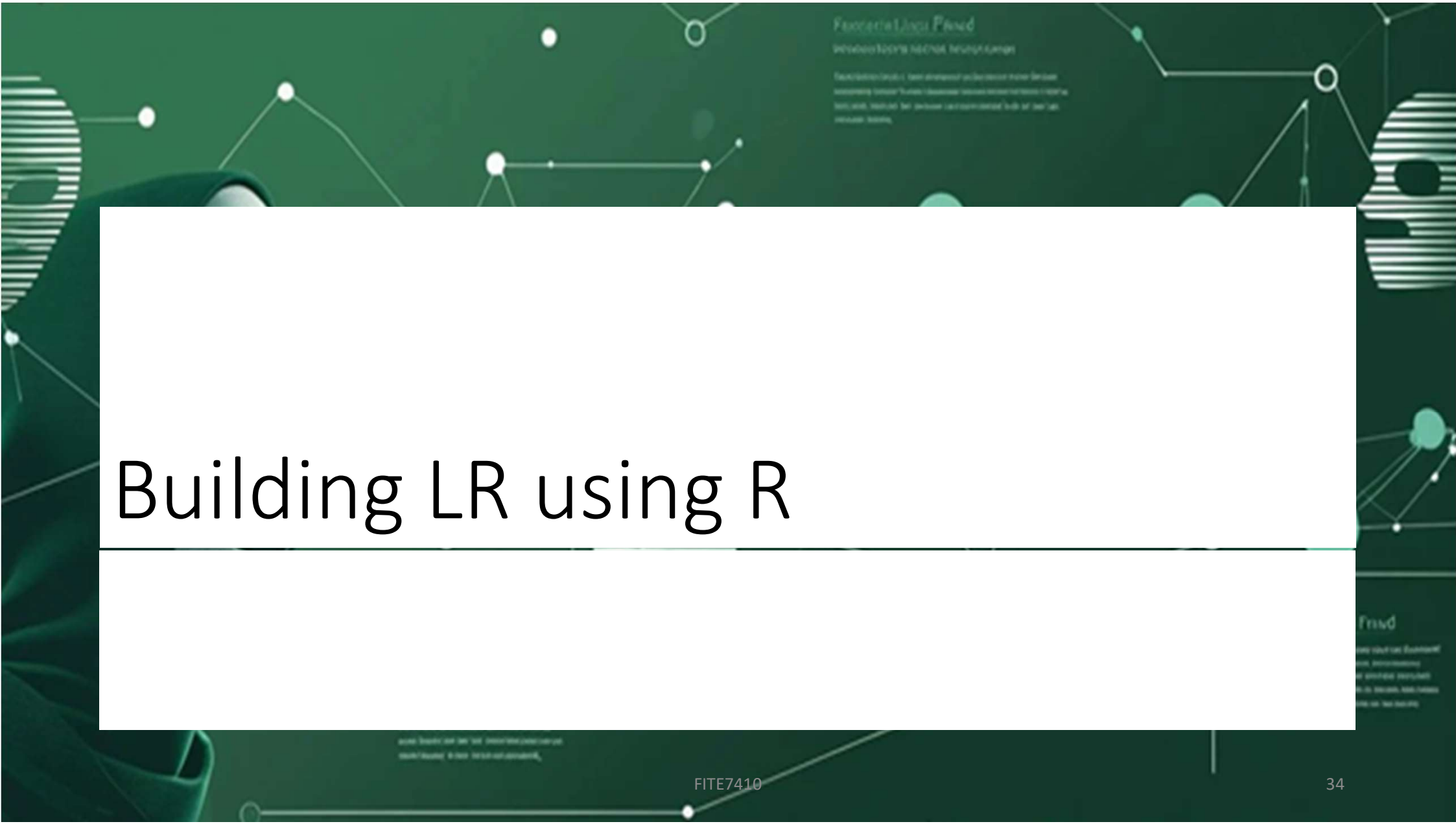
- SAME as Linear Regression

## Disadvantages

- SAME as Linear Regression

# 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. |



# Building LR using R

# Example R code for LR

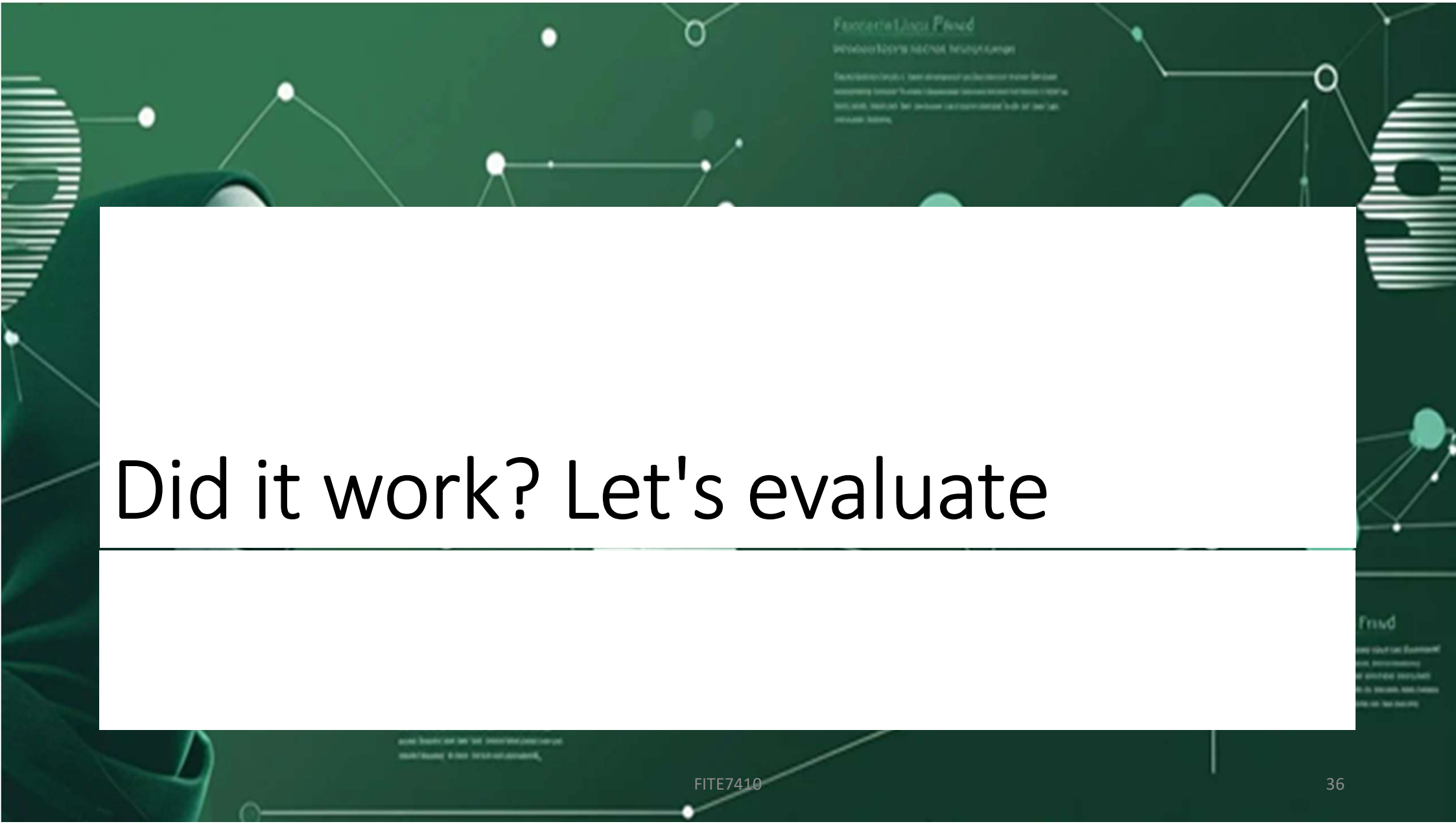
```r
# Using Logistic Regression as the modelling method
fit.lm <- glm(Class ~., data = train_rose, family = binomial(link="logit"))

# Show the LR result
summary(fit.lm)
```

glm = function for logistic regression

Specifies to run logistic regression

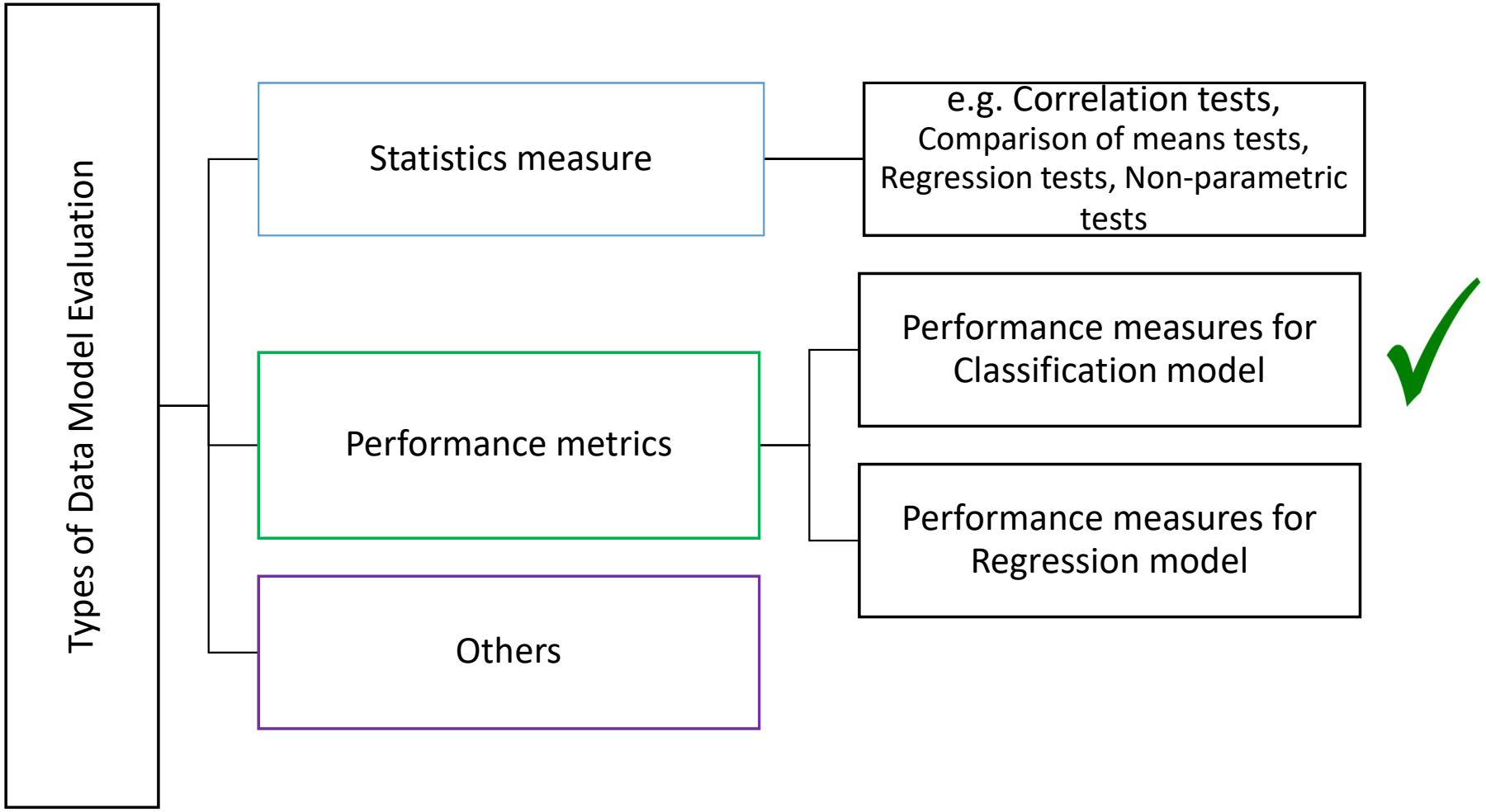
summary = function to show result of the model

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.302e+00  3.496e-02 -94.448  < 2e-16 ***
Time        -9.597e-06  2.436e-07 -39.389  < 2e-16 ***
V1           6.856e-01  1.557e-02  44.039  < 2e-16 ***
V2           5.769e-01  1.991e-02  28.979  < 2e-16 ***
V3           2.116e-01  1.207e-02  17.531  < 2e-16 ***
V4           7.148e-01  7.193e-03  99.384  < 2e-16 ***
                        ⋮
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 629920  on 454391  degrees of freedom
Residual deviance: 124265  on 454361  degrees of freedom
AIC: 124327

Number of Fisher Scoring iterations: 14
```



# Did it work? Let's evaluate

# Types of Data Model Evaluation



# Confusion Matrix

- The Confusion Matrix – an example

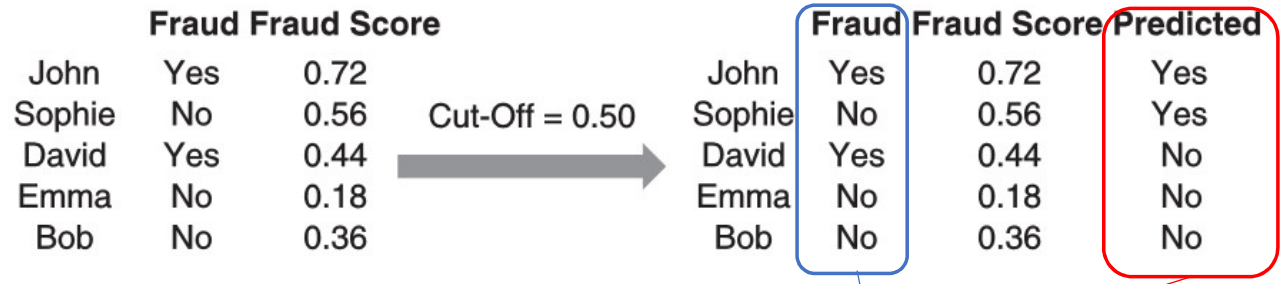



**Figure 4.37** Calculating Predictions Using a Cut-Off

**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) |

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

**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 |

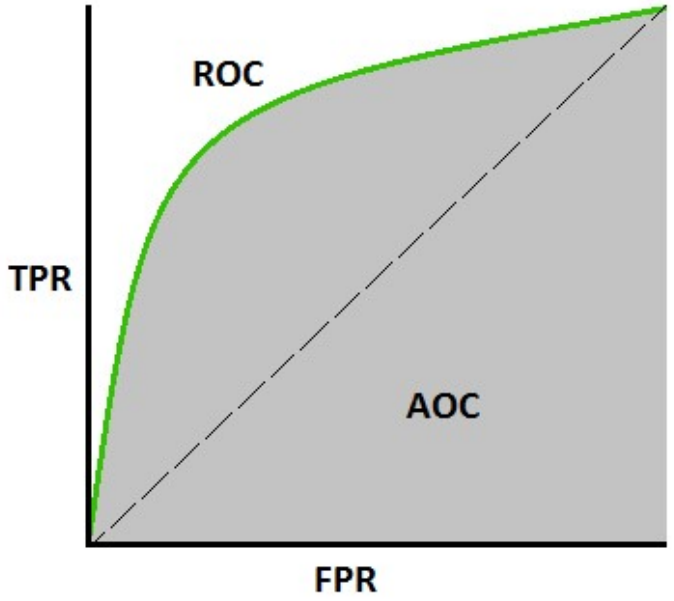
**Accuracy = (2+87) / 100 = 89%**
**Recall = (2)/(2+8)=20%**
**Precision = (2)/(2+3)=40%**
**F1 score = 2x20%x40%/(20%+40%) = 27%**

How to interpret these 4 measures?

For fraud detection models, which performance metrics is most useful performance measure?

# 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



# 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 : Example

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

| | Fraud | Fraud Score | | | Fraud | Fraud Score | Predicted |
|---|---|---|---|---|---|---|---|
| John | Yes | 0.72 | | John | Yes | 0.72 | Yes |
| Sophie | No | 0.56 | Cut-Off = 0.50 | Sophie | No | 0.56 | Yes |
| David | Yes | 0.44 | → | David | Yes | 0.44 | No |
| Emma | No | 0.18 | | Emma | No | 0.18 | No |
| Bob | No | 0.36 | | 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

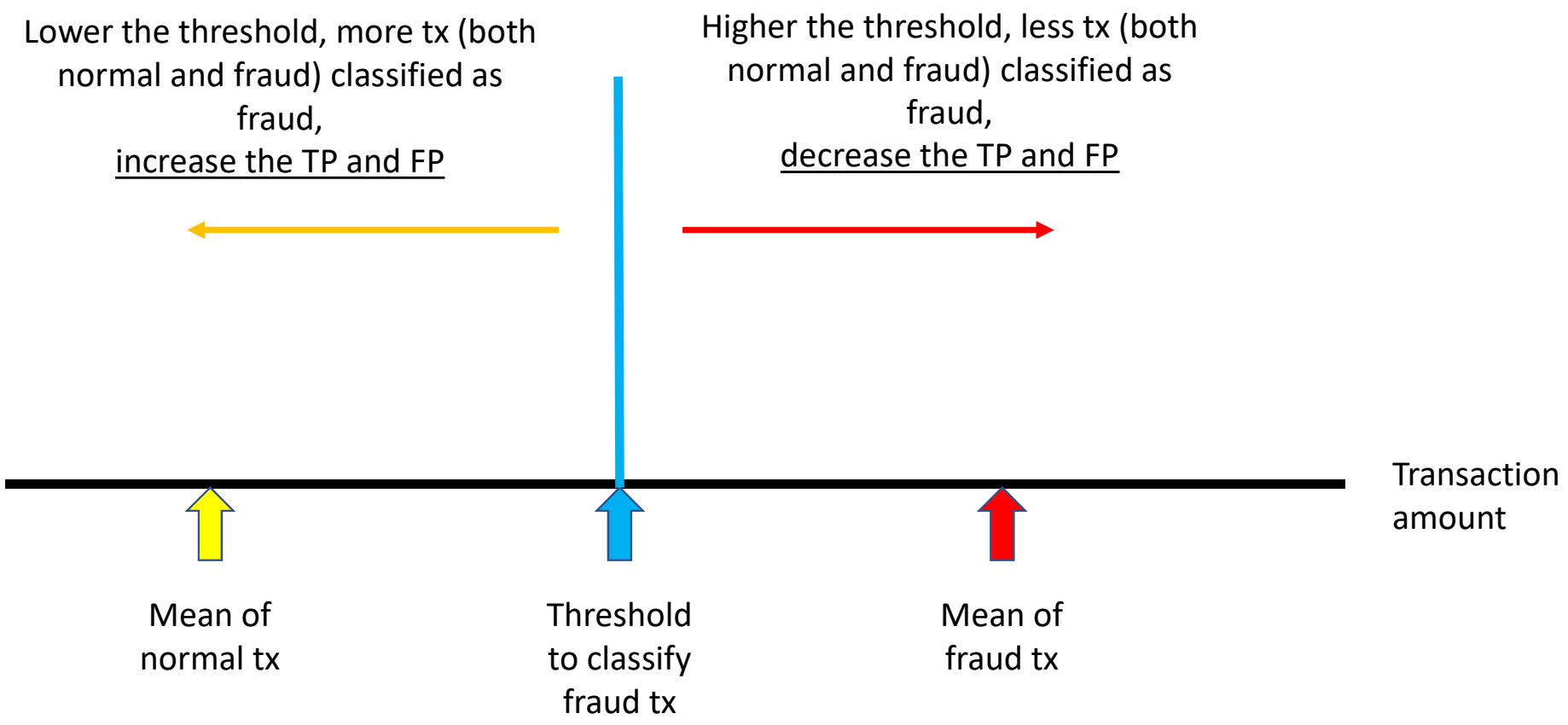
**Table 4.6** Table for ROC Analysis

| 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%



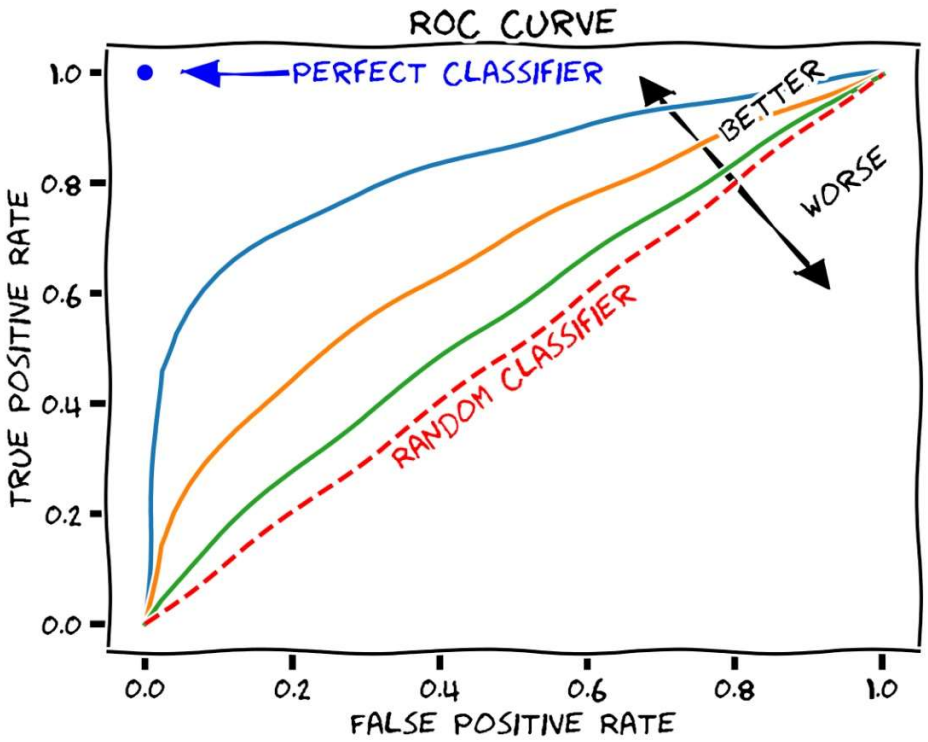
Source: Fraud Analytics using Descriptive, Predictive, and Social Network Techniques, 1st ed, John Wiley & Sons Inc.




What is the optimal threshold?

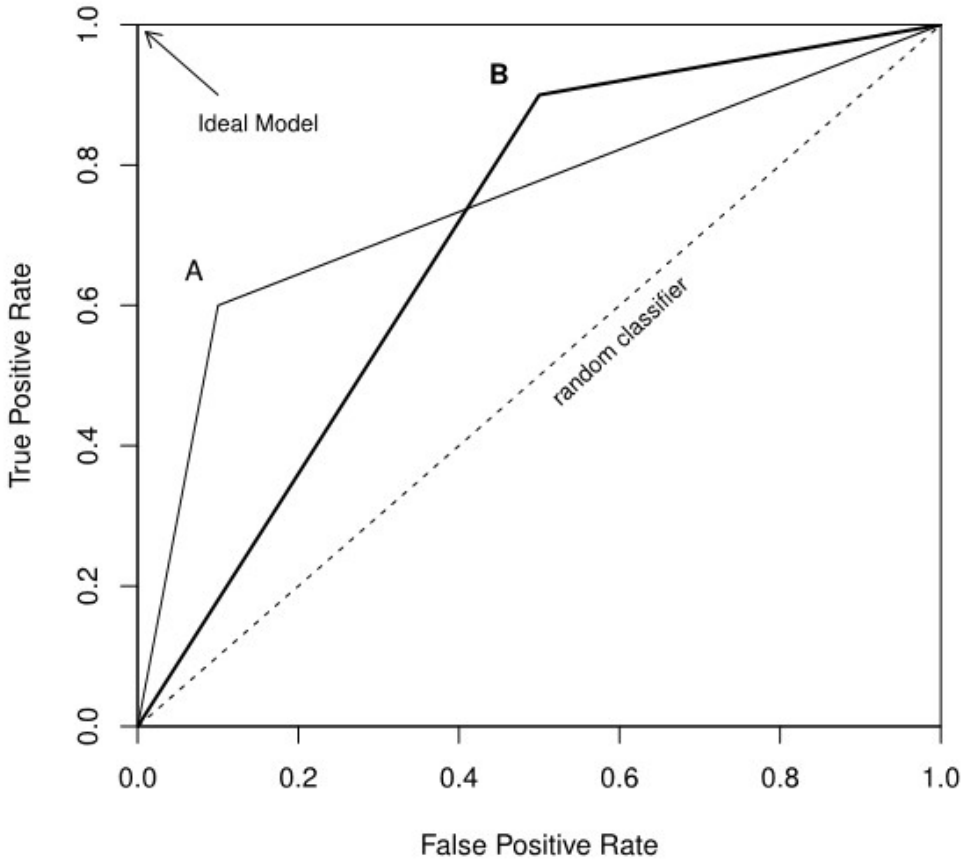
# ROC curve (Receiver Operating Characteristic curve)

- Change the threshold, show the TPR vs FPR
- ROC curve: a graph showing the performance of a classification model for all classification thresholds
- *Lowering the classification threshold classifies more items as positive*
  - *increasing both False Positives and True Positives.*



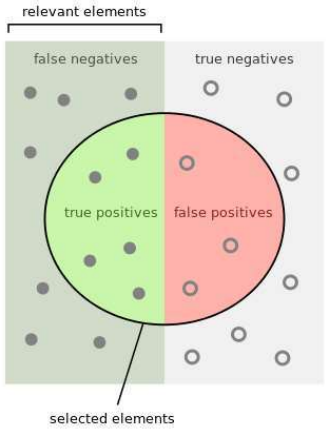

Picture from https://en.wikipedia.org/wiki/Receiver_operating_characteristic

# 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

# Another view – Precision vs Recall



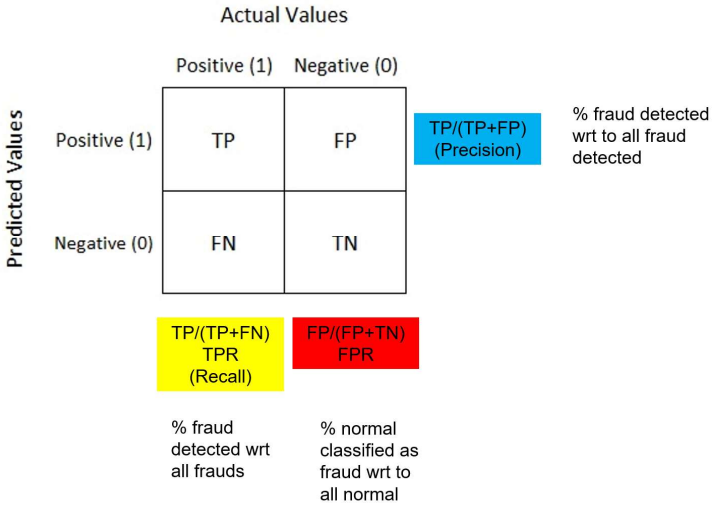

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

Precision measures the relevancy of obtained results

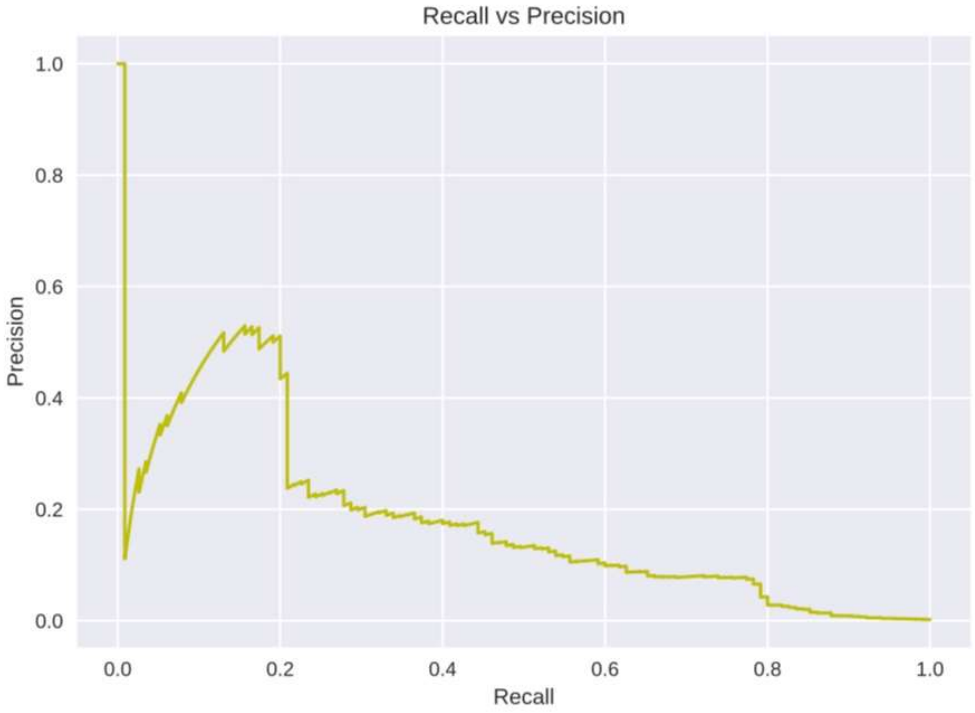
$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Recall measures how many relevant results are returned

- Good to have both Precision and Recall values equal to 1
- High recall low precision: classify many fraud transactions, most of them are normal transactions (high false positives, low relevancy)
- High precision low recall: classify few fraud transactions (with high relevancy), many fraud transactions cannot be found



# Sample evaluation result – Precision-Recall Curve (PRC)




- A high area under the curve represents both high recall and high precision
  - High precision means to false positive rate
  - High recall means low false negative rate
- High scores for both precision and recall means the classifier returns accurate results (high precision) and finds majority of all fraudulent transactions (high recall)

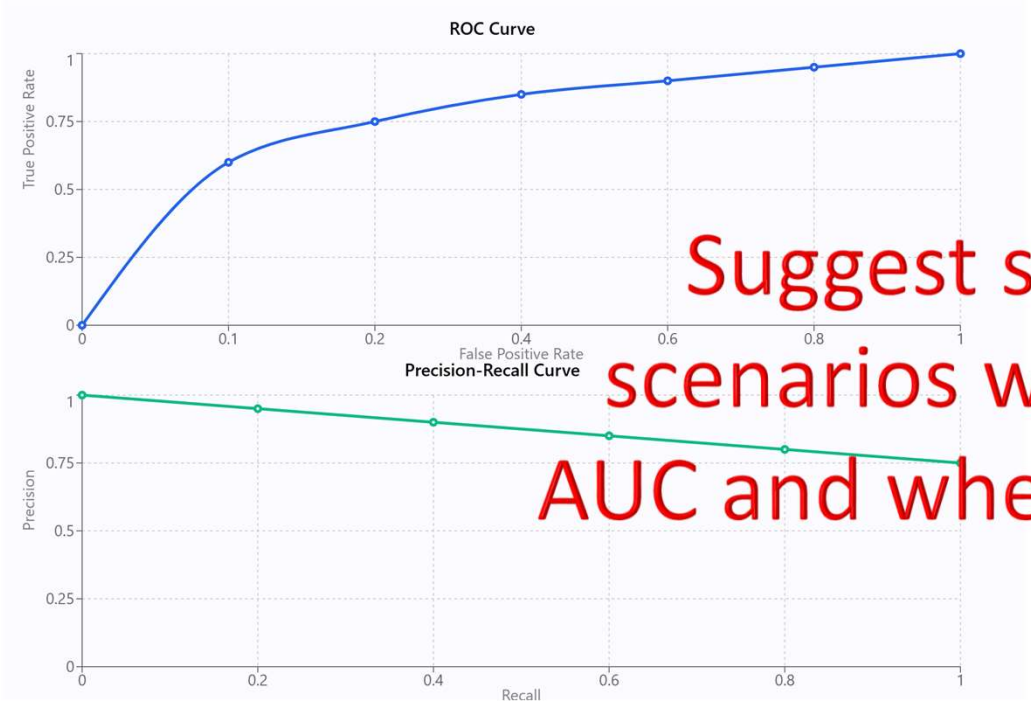
# ROC AUC vs PR AUC

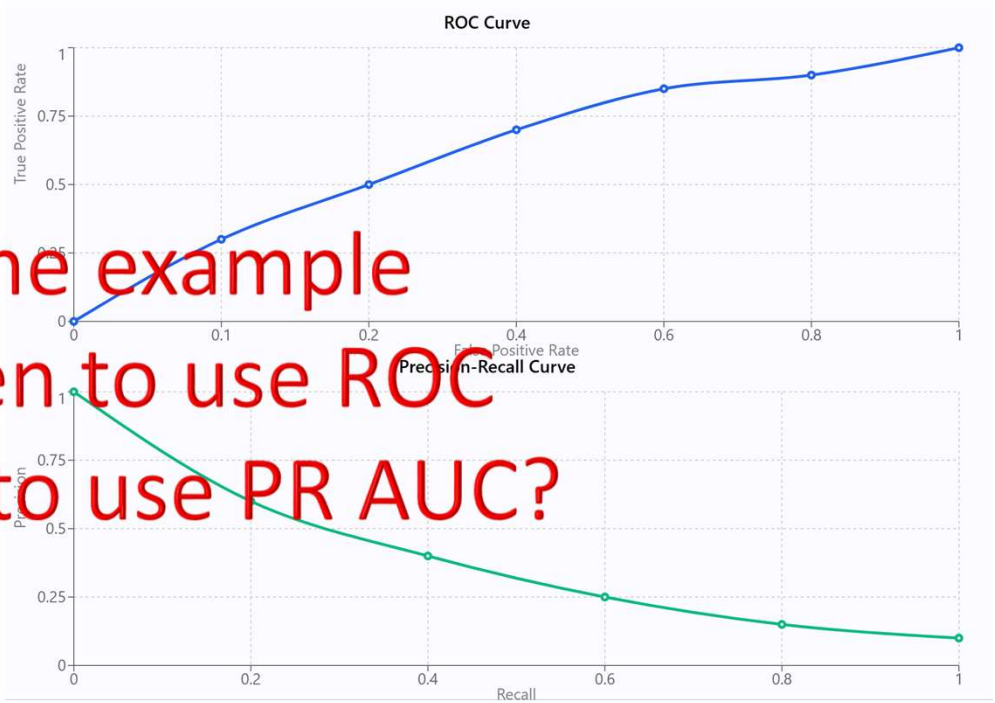
| | **ROC AUC (Receiver Operating Characteristic - Area Under Curve)** | **PR AUC (Precision-Recall - Area Under Curve)** |
|---|---|---|
| **Curve plots** | **True Positive Rate (TPR)** against **False Positive Rate (FPR)** at various threshold settings | **Precision** against **Recall** at various thresholds |
| **AUC (Area Under the Curve)** | Gives a single scalar value between 0 and 1:<br>1 = perfect classifier<br>0.5 = random guessing | Summarizes the trade-off between precision and recall across thresholds |
| **Key Concepts** | TPR = TP / (TP + FN)<br>FPR = FP / (FP + TN) | Precision = TP / (TP + FP)<br>Recall = TP / (TP + FN) |

# ROC AUC vs PR AUC

| | **ROC AUC (Receiver Operating Characteristic - Area Under Curve)** | **PR AUC (Precision-Recall - Area Under Curve)** |
|---|---|---|
| **Focus** | Separating classes (true vs. false positives) | Accuracy of positive predictions |
| **Best for** | Balanced datasets | Imbalanced datasets |
| **Insensitive to** | Class imbalance (can be misleading in imbalanced data) | More sensitive to class imbalance |
| **Interpretation** | Probability that a randomly chosen **positive** instance is ranked **higher** than a randomly chosen **negative** instance<br>(Asking the question: How well can the model distinguish between classes overall?) | The **average precision** achieved over all recall levels — i.e., **how accurate** the positive predictions are when the model is confident<br>(Asking the question: When the model says something is positive, how trustworthy is it?) |

# Balanced dataset vs Imbalanced dataset




Suggest some example scenarios when to use ROC AUC and when to use PR AUC?

**Which one is the "Imbalanced dataset"?**

# Sample R code

```r
# For the Confusion Matrix, we need the final 0/1 class predictions.
# This uses the fixed 0.5 threshold
lr_pred_class <- ifelse(predict(fit.lm, newdata = test, type = 'response') > 0.5, 1, 0)
```

Threshold

LR model

Testing dataset

```r
# The confusionMatrix() function from the 'caret' package provides a detailed table
# including the matrix, accuracy, precision, recall, and F1-score.
# Note: Ensure both predicted classes and actual classes are factors.
# We set `positive = "1"` to get metrics for the "fraud" class.

conf_matrix <- confusionMatrix(
  data = as.factor(lr_pred_class),
  reference = as.factor(test$Class),
  positive = "1"
)

# Print the full confusion matrix and associated metrics
print(conf_matrix)
```

Prediction value

Actual value

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 56178    12
         1   685    86

               Accuracy : 0.9878
                 95% CI : (0.9868, 0.9887)
    No Information Rate : 0.9983
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1955

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.87755
            Specificity : 0.98795
         Pos Pred Value : 0.11154
         Neg Pred Value : 0.99979
             Prevalence : 0.00172
         Detection Rate : 0.00151
   Detection Prevalence : 0.01354
      Balanced Accuracy : 0.93275

       'Positive' Class : 1
```

# Plotting ROC and Precision-Recall Curve

- Step 1

```
library(pROC)     # For the ROC curve and AUC
library(PRROC)    # For the Precision-Recall curve and AUC
```

- Step 2

```
# For ROC and PR curves, we need the raw predicted probabilities.
lr_pred_prob <- predict(fit.lm, newdata = test, type = 'response')
```

# Plotting ROC and Precision-Recall Curve

- Step 3 – ROC AUC

```r
# The roc() function from the 'pROC' package takes the TRUE labels and the PREDICTED PROBABILITIES.
roc_obj <- roc(test$Class, lr_pred_prob)

# Plot the ROC curve
plot(roc_obj, main = "ROC Curve for Logistic Regression", print.auc = TRUE)

# You can also get the AUC value directly
roc_auc_value <- auc(roc_obj)
print(paste("Area Under ROC Curve (AUC):", roc_auc_value))
```

Predicted probabilities

# Plotting ROC and Precision-Recall Curve

- Step 4a – PR AUC

```
# The PR curve is more informative for imbalanced datasets like fraud detection.
# We use the 'PRROC' package.
# It requires the predicted probabilities and the true class labels (as numeric).

# 1. Sanitize the true labels
# Convert from factor to character, then to numeric to avoid conversion errors.
# Coalesce any potential NAs to 0 (assuming 0 is the non-fraud class).
clean_labels <- as.numeric(as.character(true_labels))
clean_labels[is.na(clean_labels)] <- 0

# 2. Sanitize the predicted scores
# It's less common for `predict` to produce NAs, but it's good practice to handle them.
# We'll replace any NAs in predictions with 0, a safe non-fraud score.
clean_scores <- pred_scores
clean_scores[is.na(clean_scores)] <- 0
```

Convert to “numeric” type

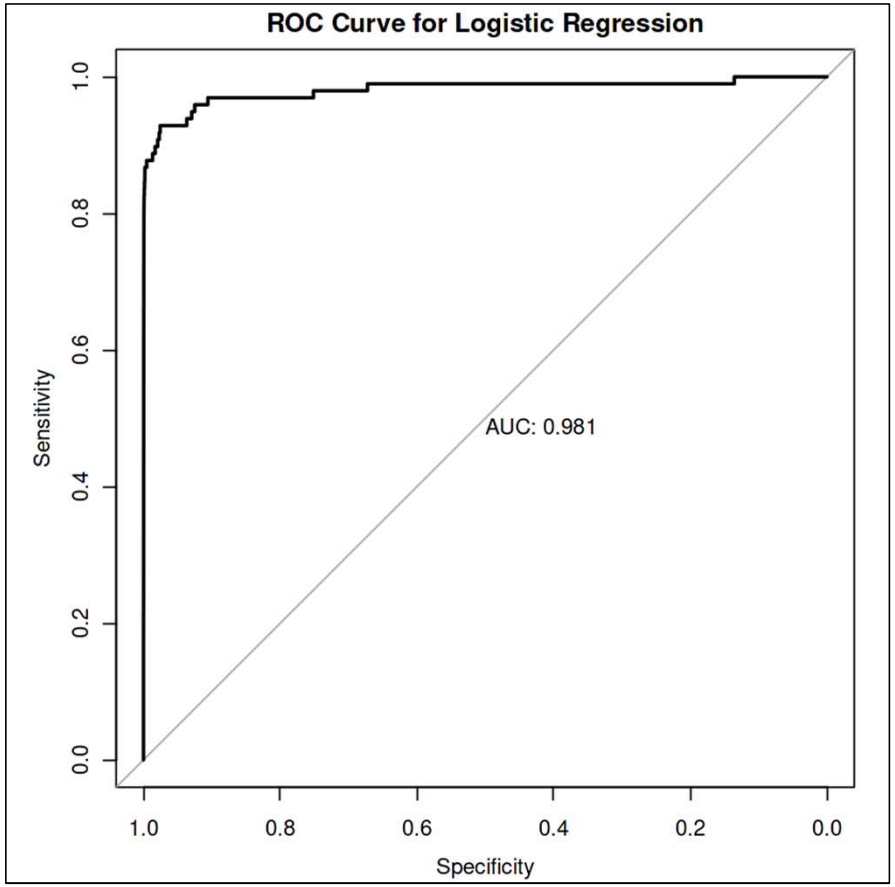
To handle any missing value

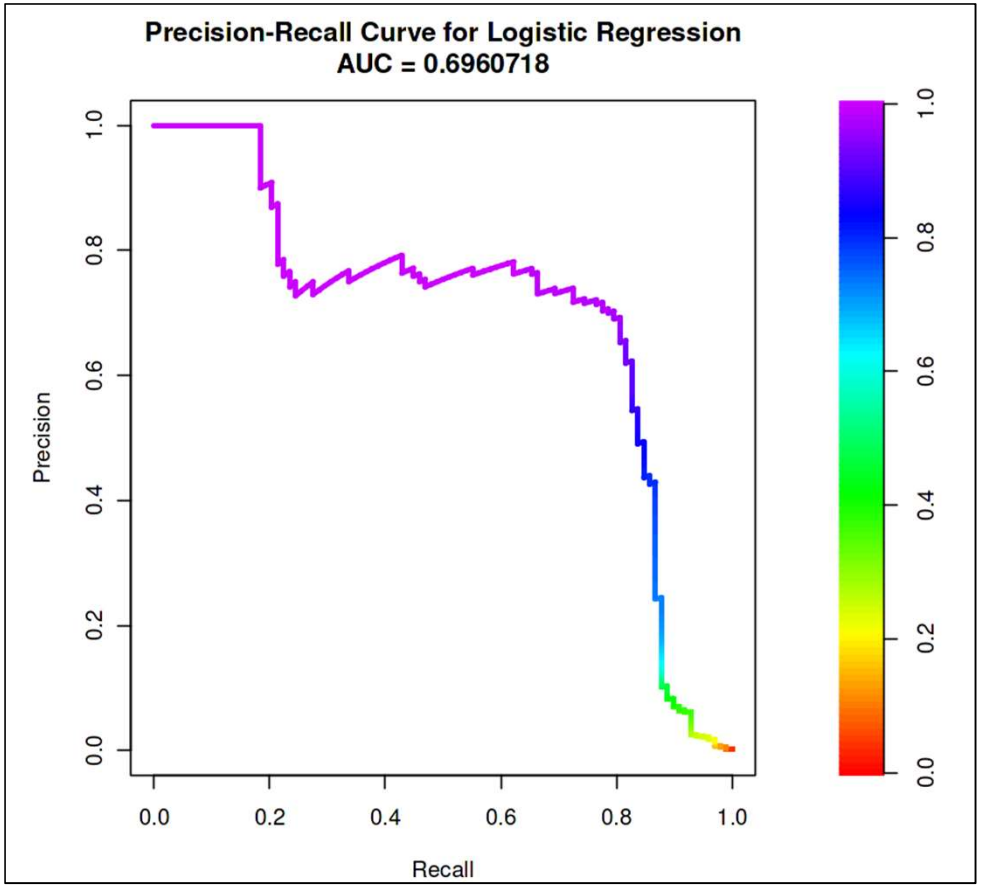
# Plotting ROC and Precision-Recall Curve

- Step 4b – PR AUC (plot the graph using transformed data)

```
# Run the function with the sanitized vectors
pr_obj <- pr.curve(
  scores.class0 = clean_scores,     # Use the clean predicted scores
  weights.class0 = clean_labels,    # Use the clean numeric labels
  curve = TRUE
)

# Plot the curve and print the AUC
plot(pr_obj, main = "Precision-Recall Curve for Logistic Regression")
print(pr_obj$auc.integral)
```

# Plotting ROC and Precision-Recall Curve



# Case studies : Financial Statement Fraud cases

# 2 Case Studies that applied LR for fraud detection

- Case study #1 : Romanian Study
  - *Sabău, A.-I., Gherai, R. S., & Todea, A. (2021). A statistical model of fraud risk in financial statements. Case for companies listed on the Bucharest Stock Exchange. Risks, 9(12), 222.*
- Case study #2 : Chinese Study
  - *Guan, R., Lin, T., & Liu, X. (2022). Financial fraud identification of the companies based on multi-relationship analysis of financial statements. Prague Economic Papers, 31(2), 211–231.*

# Case Study #1 Background – The Romanian Market

- Focus:
  - Examines companies listed on the Bucharest Stock Exchange, an emerging Eastern European market.
- Approach:
  - Applies the classic Beneish M-Score model, which uses 8 standard financial ratios.
- Objective:
  - Research question: “What are the financial indicators that most strongly discriminate the two states: fraudulent financial statements and non-fraudulent financial statements?”
- Dataset:
  - A sample of 66 companies analyzed over a 5-year period (2015-2019).
  - An average Beneish score was made for each company. The resulting values were reported at the reference level of “-2.22”. The scores higher than the reference point were included in the “FRAUD” area, and those with a lower score “NON-FRAUD”. Of the companies, 44 belong to the “FRAUD” area and 22 “NON-FRAUD”.

# Case Study #2 Background – The Chinese Market

- Focus:
  - A study on corporate fraud detection within publicly listed companies in China.
- Approach:
  - Develops a custom logistic regression model using a broad set of 17 indicators (12 financial, e.g. X1= (Accounts receivable growth rate )/(Operating revenue growth rate) ; and 5 non-financial, e.g. Whether the general manager is concurrently chairman).
- Objective:
  - To build a predictive model capable of identifying fraudulent firms based on this diverse set of indicators.
- Dataset:
  - A balanced sample of 106 firms, comprising 53 known fraudulent companies and 53 non-fraudulent counterparts.
  - The study selects the enterprises with financial violations in the China Stock Market and Accounting Research (2022) database (CSMAR) as the fraud sample for 2017–2020.

# Summary

| | Romanian Study (SabÄƒfu et al.) | Chinese Study (Guan et al.) |
|---|---|---|
| Dataset | 66 companies over 5 years from a single emerging market. 44 belong to the “FRAUD” with Beneish score over “-2.22”. | Balanced sample of 106 companies (53 known fraud, 53 non-fraud). |
| Variables | Standard 8 financial ratios from the established Beneish M-Score model. | Custom-selected 17 indicators, including 5 non-financial metrics (e.g., Whether the internal control is valid). |
| Modeling | Standard application of Logistic Regression on Beneish scores. | Logistic Regression with extensive feature engineering and selection. |

# Logistic Regression Result (1)

- Case #1 - Romanian Study

**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 |

**Table 4.** Multivariate logistic regression.

| Variables | B | Exp(B) | *p*-Value |
|---|---|---|---|
| GMI | 4.234 | 68.964 | 0.028 |
| AQI | 1.971 | 7.175 | 0.149 |
| DEPI | 0.755 | 2.128 | 0.129 |
| TATA | 23.862 | $2.3 \times 10^{10}$ | 0.004 |
| Constant | −7.095 | 0.001 | 0.011 |

Source: authors' calculations in SPSS 24.

# Logistic Regression Result (1)

- Case #2 Chinese Study - Logistic Regression result

Tab. 3 – The outputs of logistic regression

| Variable | Coeff. | St. Er. | Variable | Coeff. | St. Er. | Variable | Coeff. |
|---|---|---|---|---|---|---|---|
| X1 | -0.01 | 0.06 | 0.93 | X10 | -1.06 | 0.41 | 0.01 |
| X2 | 0.65 | 1.47 | 0.66 | X11 | 0.00 | 0.13 | 1.00 |
| X3 | -0.10 | 0.05 | 0.05 | X12 | -0.44 | 0.34 | 0.20 |
| X4 | 0.02 | 0.04 | 0.60 | C1 | 0.03 | 0.02 | 0.17 |
| X5 | 24.03 | 13.15 | 0.07 | C2 | -0.24 | 0.60 | 0.69 |
| X6 | 0.04 | 0.06 | 0.54 | C3 | 10.20 | 5.42 | 0.06 |
| X7 | 0.13 | 0.08 | 0.10 | C4 | 1.98 | 2.04 | 0.33 |
| X8 | -3.19 | 7.59 | 0.67 | C5 | 0.41 | 1.77 | 0.82 |
| X9 | 0.27 | 0.30 | 0.36 | const. | -7.80 | 3.38 | 0.02 |
| -2 Log likelihood | | | | | | | 82.506a |
| Cox&Snell R2 | | | | | | | 0.46 |
| Nagelkerke R2 | | | | | | | 0.61 |

Note: Coeff. – Regression Coefficient; St. Er. – Standard Error; Consp. – Conspicuousness

**How to interpret this LR result?**

# Classification performance results

Case study #1
Romanian Study

**Table 5.** Classification Table for the multivariate binary regression.

| Observed | | Predicted: Fraud: Non-Fraud | Predicted: Fraud: Fraud | Percentage Correct |
|---|---|---|---|---|
| Fraud | Non-fraud | 20 | 2 | 90.9 |
| | Fraud | 2 | 42 | 95.5 |
| Overall Percentage | | | | 93.9 |

Source: authors' calculations in SPSS 24.

Case study #2
Chinese Study

Tab. 6 – The results of overall prediction accuracy for robust model (13)

| Classify | | Engage in Embezzlement: 0 | Engage in Embezzlement: 1 | Precision % |
|---|---|---|---|---|
| Engage in Embezzlement | 0 | 48 | 2 | 96 |
| | 1 | 6 | 44 | 88 |
| Overall Percentage | | | | 92 |

**How to interpret this performance metrics result?**

# Which LR model you would deploy to use?

- A. Romanian Study
- B. Chinese Study
- C. None of the studies

Any methodological issues with these 2 studies?

# Which performance metrics for LR model is more useful for fraud detection?

- A. R-square / AIC / BIC
- B. Precision
- C. Recall
- D. Accuracy
- E. None of the above

What is your choice? Why?

# Explanation vs. Prediction

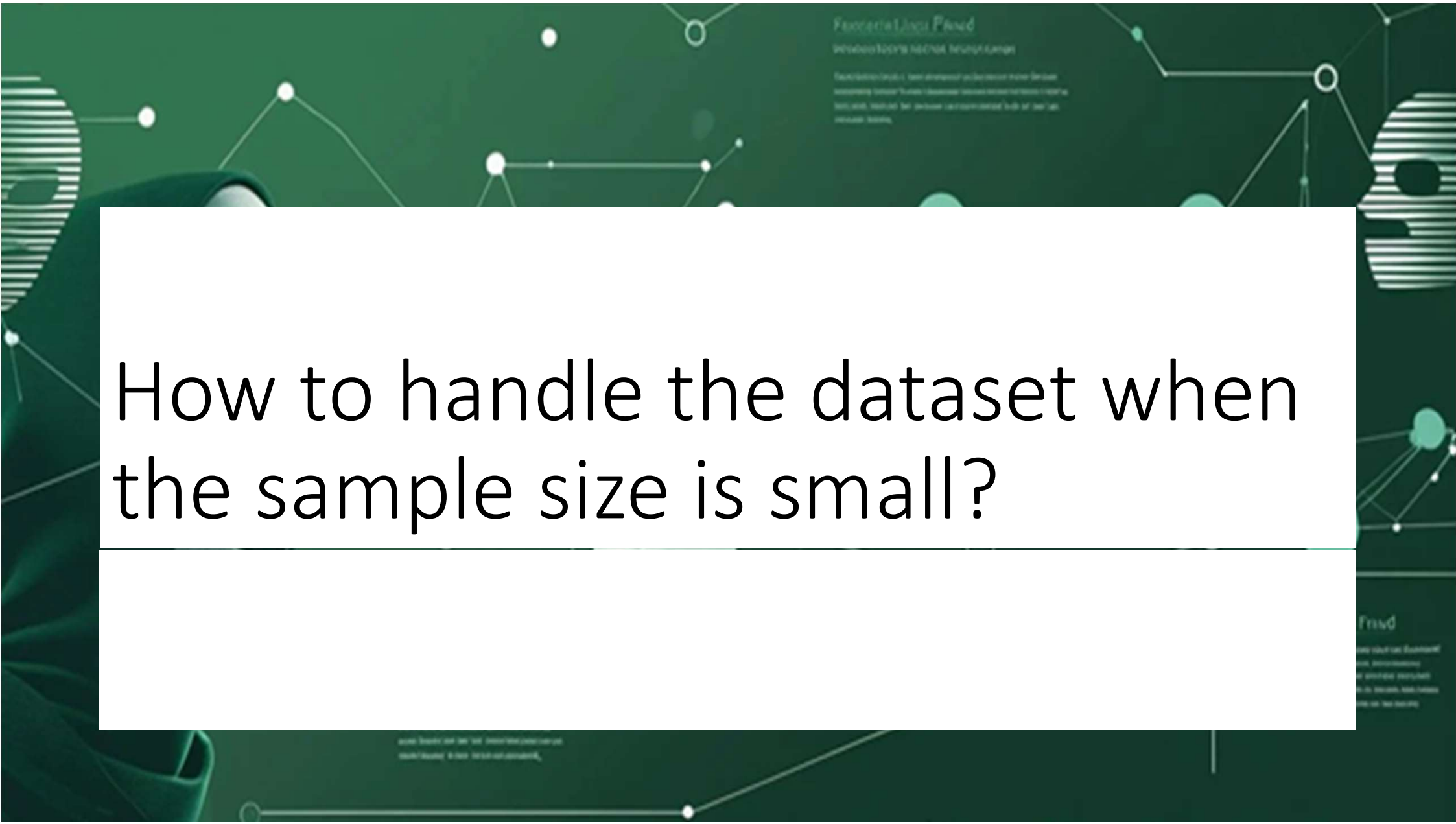
- The confusion arises because logistic regression can be used for two distinct purposes:
  - Explanation (Inference):
    - The goal is to understand the relationship between variables.
    - You want to know which factors (e.g., DSRI, GMI) significantly influence the probability of fraud. Here, model fit statistics and the significance of coefficients (p-values) are important. You are asking, "Does my model accurately describe the data I have?"
  - Prediction (Classification):
    - The goal is to correctly classify new, unseen cases.
    - You want to build a system that reliably flags a transaction as fraudulent or not. Here, classification metrics are what matter. You are asking, "How well does my model perform its assigned task in the real world?

# A Tiered Evaluation Strategy for Fraud Models

- Tier 1: Business-Critical Classification Metrics (Primary Focus)
  - Recall: Is our fraud capture rate high enough to protect the business?
  - Precision: Is our false positive rate low enough to be operationally sustainable?
  - Precision-Recall (PR) Curve: This is the single most important visualization for imbalanced classification. It shows the trade-off between precision and recall across all possible thresholds. The Area Under the PR Curve (PR-AUC) is a superior metric to the standard ROC-AUC for imbalanced data.
  - F1-Score: This provides a single number that balances precision and recall, which is useful for comparing models at a glance.

# A Tiered Evaluation Strategy for Fraud Models

- Tier 2: Model-Level Goodness-of-Fit (Secondary Focus)
  - R-squared: Use these *only* to compare the relative fit of different models on the same data. For example, if you add a new variable and the R-squared increases, it suggests the new variable added some explanatory power. However, this is meaningless if the Tier 1 classification metrics did not also improve.
  - AIC/BIC: Similar to R-squared, these are useful for model selection (e.g., comparing a model with 8 variables to one with 17), as they penalize models for unnecessary complexity.



# How to handle the dataset when the sample size is small?

# K-fold cross-validation

- Cross-validation can be applied when the sample size is very small, e.g. less than 1000 observations


Pic Source: Fedotenkova, Mariia. (2016). Extraction of multivariate components in brain signals obtained during general anesthesia.

A schematic illustration of K-fold cross-validation for K = 5.

- Original dataset (shown in dark green) is randomly partitioned into K disjoint sets (shown in light green).
- Then K − 1 parts are used for training a model (shown in blue) and remaining part is used for evaluation (shown in orange).
- This process is repeated K times for all possible choices of the test set, producing test errors.
- The final performance is reported by averaging the errors from each iteration.

# How to choose value of K?

- **Representative**
  - Choose a value for k so that each train/test group samples is statistically representative of the population dataset
- **k=5 or 10**
  - A commonly chosen value as this value has been found to generally result in a model estimate with low bias and a modest variance
- **k=n**
  - ‘n’ is the size of the dataset, i.e. leave-one-out cross-validation. That means each test sample is given an opportunity to be used in the hold out dataset

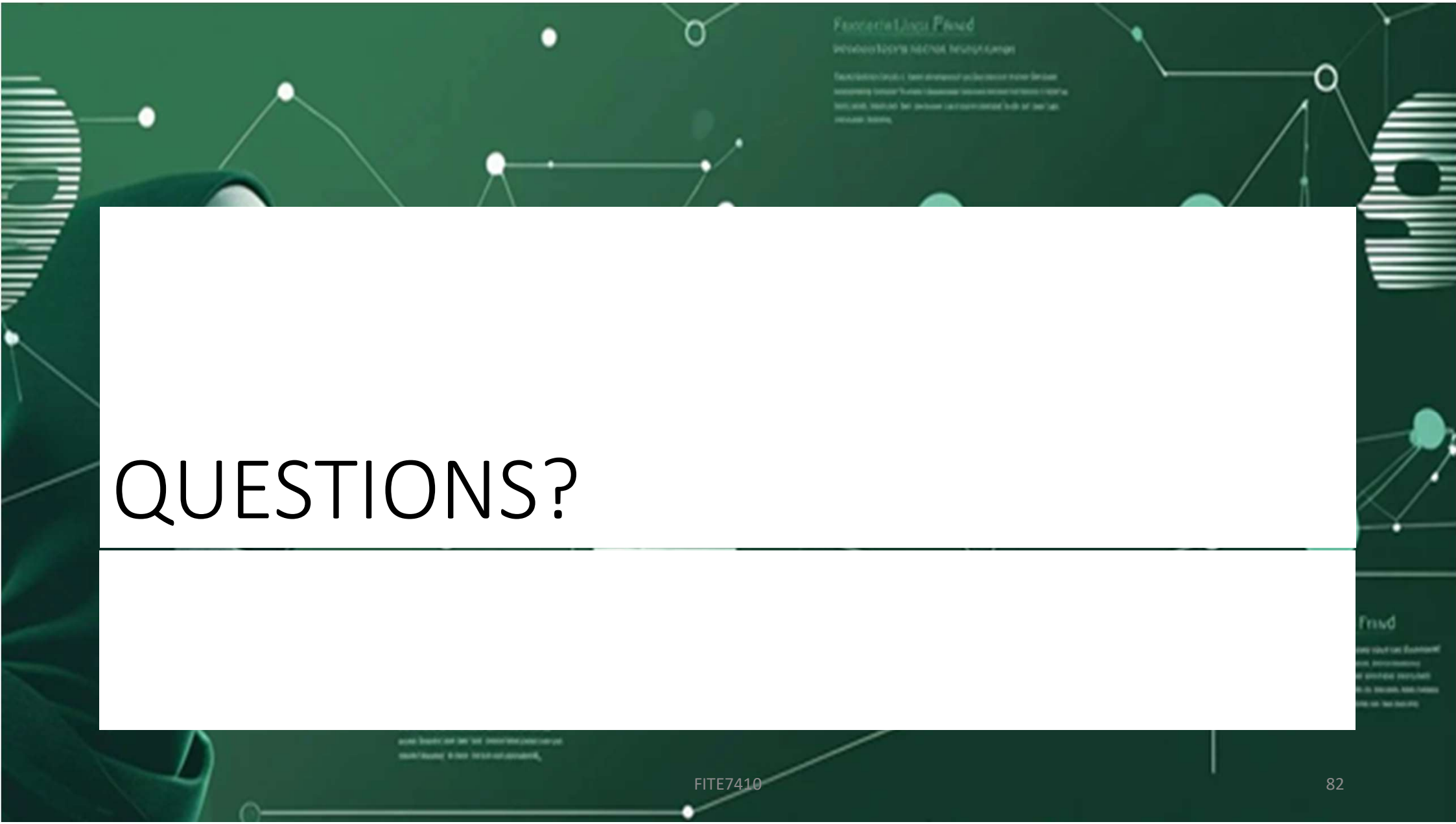
# Cross-validation

- Q : With more than one trained model, choose which model?

- A :
  - Similar to ensemble method, use voting procedure.
  - Use leave one out cross-validation and randomly select one model. Since all models differ by one observation only, the performance should be similar for all models.
  - Use all observations for training. Then, use the cross-validation performance result as the independent estimate of the model.

# References


- Bart Baesens, Veronique Van Vlasselaer, Wouter Verbeke (2015). Fraud Analytics using Descriptive, Predictive, and Social Network Techniques, 1st ed, John Wiley & Sons Inc.
- Guan, H., Li, S., Wang, Q., Lyulyov, O. & Pimonenko, T. (2022). Financial Fraud Identification of the Companies Based on the Logistic Regression Model. Journal of Competitiveness, 14(4), 155–171. https://doi.org/10.7441/joc.2022.04.09
- Leonard W. Vona (2017). Fraud Data Analytics Methodology: The Fraud Scenario Approach to Uncovering Fraud in Core Business Systems, John Wiley & Sons, Inc.
- 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.
- Sabău, A.-I., Gherai, R. S., & Todea, A. (2021). A statistical model of fraud risk in financial statements. Case for companies listed on the Bucharest Stock Exchange. *Risks*, *9*(12), 222. https://doi.org/10.3390/risks9120222



# QUESTIONS?