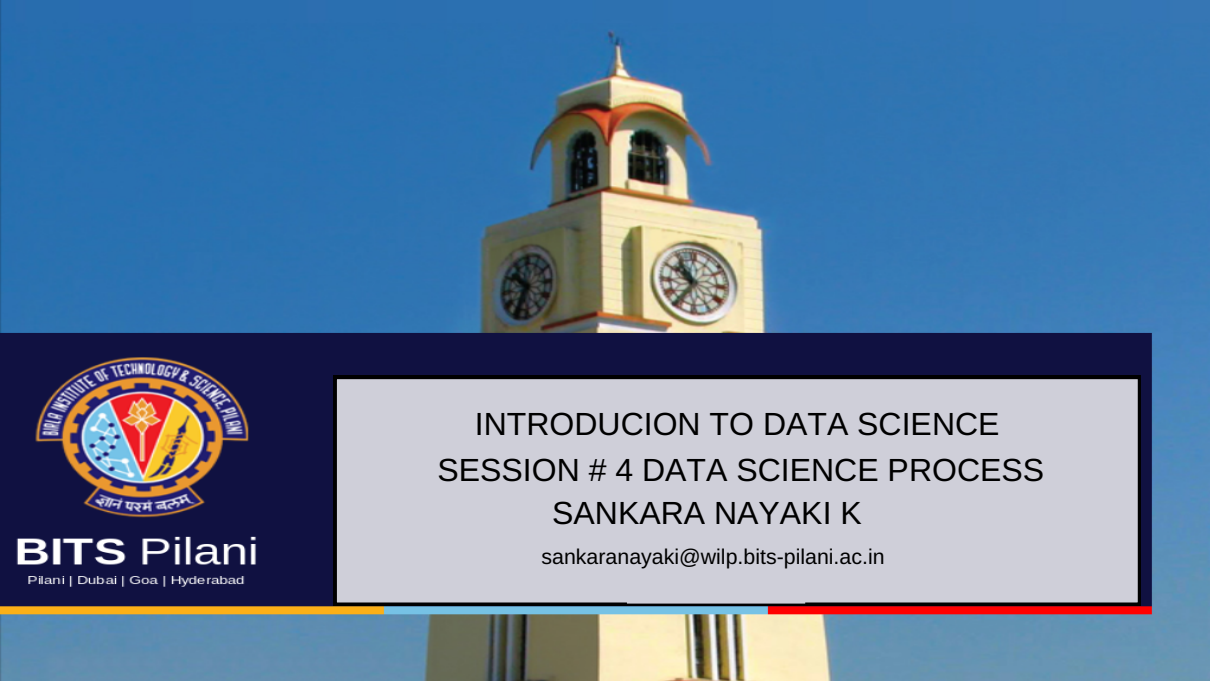


BITS Pilani
Pilani | Dubai | Goa | Hyderabad

# INTRODUCION TO DATA SCIENCE
# SESSION # 4 DATA SCIENCE PROCESS

SANKARA NAYAKI K
sankaranayaki@wilp.bits-pilani.ac.in

The instructor is gratefully acknowledging
the authors who made their course
materials freely available online.

References:


- Introducing Data Science by Cielen, Meysman and Ali
- Storytelling with Data by Cole Nussbaumer Knaflic; Wiley
- Introduction to Data Mining by Tan, Steinbach and Vipin Kumar
- The Art of Data Science by Roger D Peng and Elizabeth Matsui
- Python Data Science Handbook: Essential tools for working with data by Jake VanderPlas

# Table of Contents

# Course Handout

M1 Introduction to Data Science
M2 Data Analytics
M3 Data Science Process
M4 Data Science Teams
M5 Data & Data Models
M6 Data Wrangling & Feature Engineering
M7 Data Visualization
M8 Storytelling with Data
M9 Ethics for Data Science

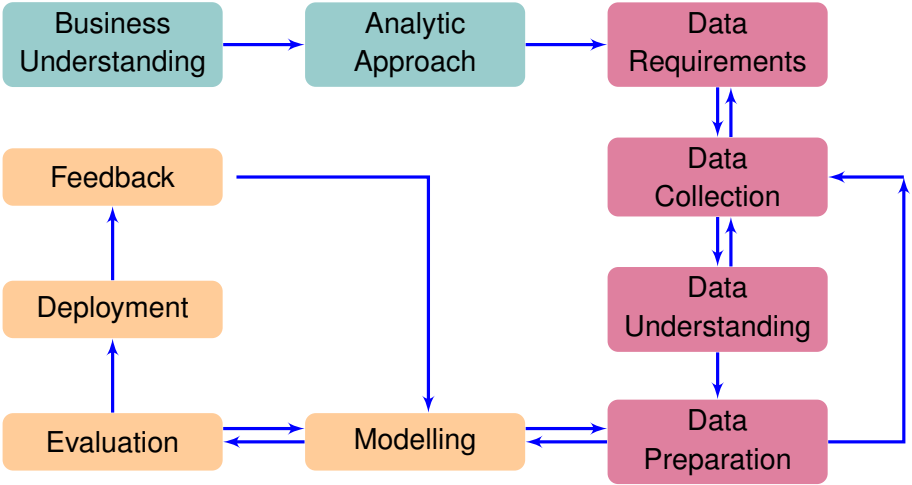
# Table of Contents

# Data Science Methodology

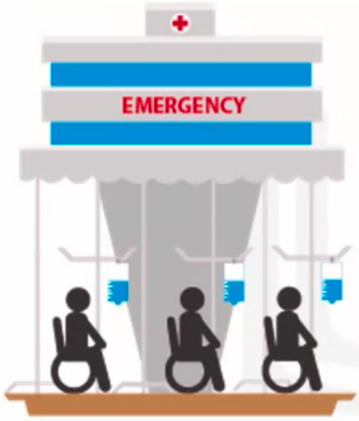
- Problem to Approach
  1. What is the problem that you are trying to solve?
  2. How can you use data to answer the questions?
- Working with Data
  3. What data do you need to answer the question?
  4. Where is the data coming from? Identify all Sources. How will you aquire it?
  5. Is the data that you collected representative of the porblem to be solved?
  6. What additional work is required to manipulate and work with the data?
- Delivering the Answer
  7. In what way can the data be visualized to get to the answer that is required?
  8. Does the model used really answer the initial question or does it need to be adjusted?
  9. Can you put the model into practice?
  10. Can you get constructive feedback into answering the question?

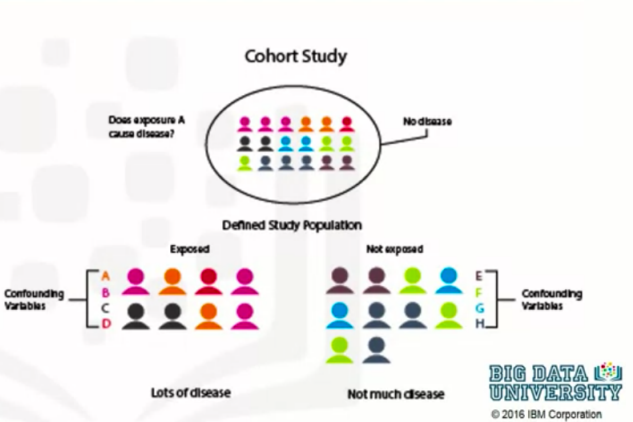
CRISP-DM approach

# Data Science Methodology



# Case Study


EMERGENCY


Cohort Study
Does exposure A cause disease?
No disease
Defined Study Population
Exposed
Not exposed
Confounding Variables
A
B
C
D
E
F
G
H
Confounding Variables
Lots of disease
Not much disease
BIG DATA UNIVERSITY
© 2016 IBM Corporation

# 1. Business Understanding

- What is the problem that you are trying to solve?
- Ask questions
- Seek clarifications
- Identify the goals
- Identify and define the objectives that support the goal.
- Get the stakeholder buy-in and support.

# Case Study - 1. Business Understanding

## What are the goals & objectives?


**Define the GOALS**

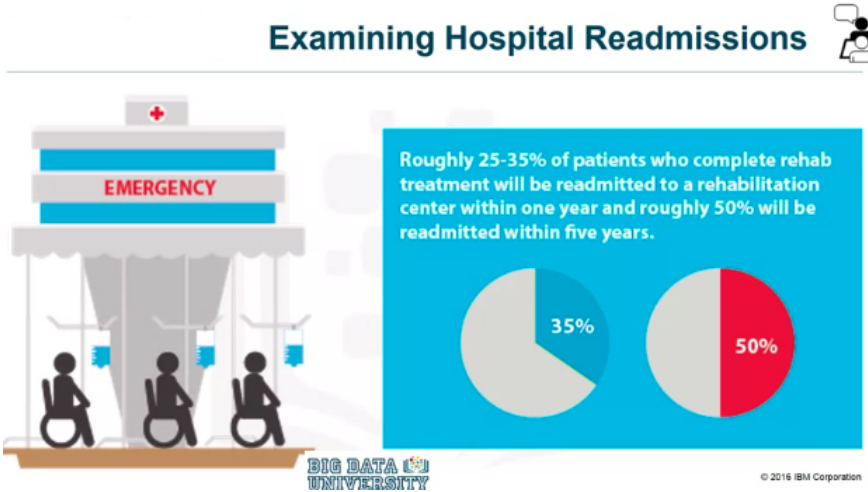
- To provide quality care without increasing cost

**Define the OBJECTIVES**

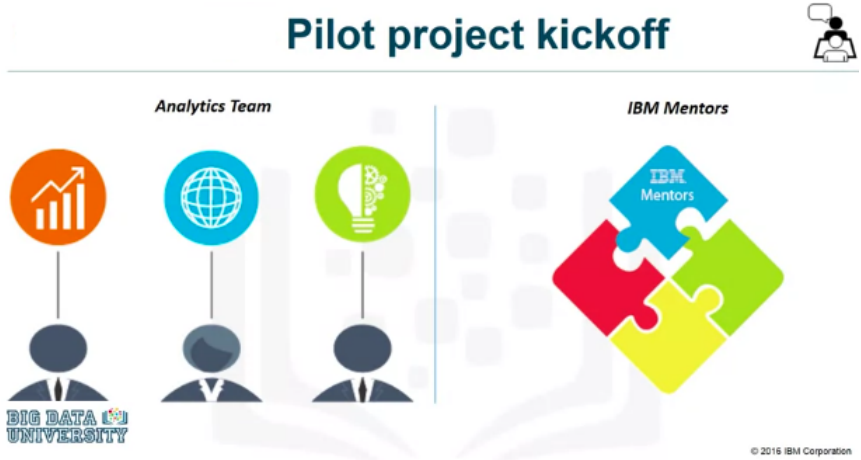
- To review the process to identify inefficiencies

BIG DATA UNIVERSITY

# Case Study - 1. Business Understanding



# Case Study - 1. Business Understanding



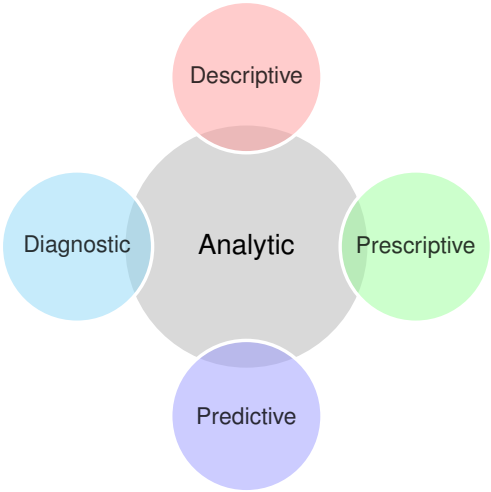
# Case Study - 1. Business Understanding

## Identifying the business requirements


1. Predict CHF readmission outcome (Y or N) for each patient
2. Predict the readmission risk for each patient
3. Understand explicitly what combination of events led to the predicted outcome for each patient
4. Easy to understand and apply to new patients to predict their readmission risk

# 2. Analytic Approach

- How can we use data to answer the questions?
- Choose Analytic approach based on the type of question.
  - Descriptive
    - ★ Current status
  - Diagnostic (Statistical Analysis)
    - ★ What happened?
    - ★ Why is this happening?
  - Predictive (Forecasting)
    - ★ What if these trends continue?
    - ★ What will happen next?
  - Prescriptive
    - ★ How do we solve it?



# 2. Analytic Approach

- Determine probabilities of an action
  - Use Predictive model
- Show relatiotionships
  - Use Descriptive model
- Requires a yes / no answer
  - Use classification model

# Case Study - 2. Analytic Approach

## Decision tree classification selected!


**Predictive model**

- To predict an outcome

**Decision tree classification**

- Categorical outcome
- Explicit "decision path" showing conditions leading to high risk
- Likelihood of classified outcome
- Easy to understand and apply

# 3. Data Requirements

- What are the data requirements?
- Identify data content, formats.
- Identify sources of data collection.

# Case Study - 3. Data Requirements

## Case Study – Selecting the cohort

- Define and select cohort
  - In-patient within health insurance provider's service area
  - Primary diagnosis of CHF in one year
  - Continuous enrollment for at least 6 months prior to primary CHF admission
  - Disqualifying conditions

# Case Study - 3. Data Requirements

## Case Study – Defining the data

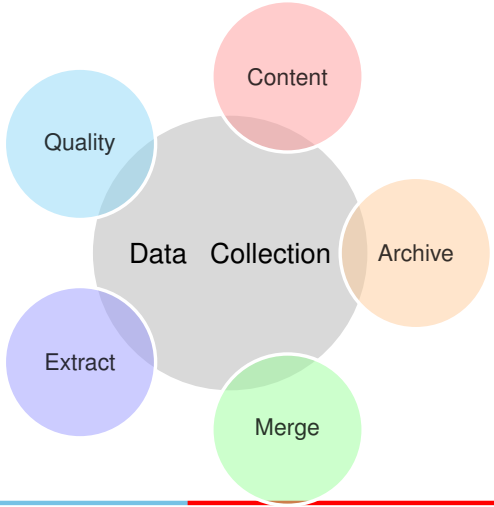
- Content, formats, representations suitable for decision tree classifier
  - One record per patient with columns representing variables (dependent variable and predictors)
  - Content covering all aspects of each patient's clinical history
    - Transactional format
    - Transformations required


BIG DATA UNIVERSITY

8

# 4. Data Collection

- What occurs during data collection?
- Assess to determine whether or not we have the data we need.
- Decide on whether more data or less data is required.
- Revise data requirements if needed.
- Assess content, quality and initial insights of data.
- Identify gaps in data.
- How to extract , merge and Archive data?



# Case Study - 4. Data Collection

## Case Study – Gathering available data

- Available data sources
  - Corporate data warehouse (single source of medical & claims, eligibility, provider and member information)
  - In-patient record system
  - Claim payment system
  - Disease management program information

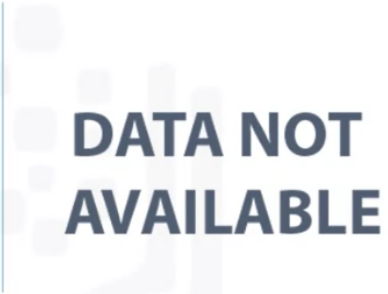
BIG DATA UNIVERSITY

7

# Case Study - 4. Data Collection

## Case Study – Deferring Inaccessible data

- Data wanted but not available
  - Pharmaceutical records
  - Decided to defer



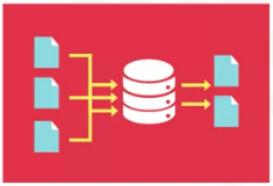

BIG DATA UNIVERSITY

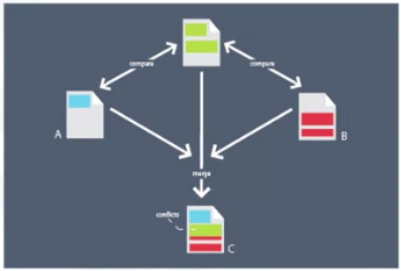
8

# Case Study - 4. Data Collection

## Case Study – Merging data

- Eliminate redundant data




BIG DATA UNIVERSITY

9

# 5. Data Understanding

- Is data you collected representative of the problem to be solved?
- Descriptitive statistics
  - Univariates statistics
  - Pairwise correlation
  - Histograms
- Assert data quality
  - Missing value
  - Invalid data
  - Misleading data

# Case Study - 5. Data Understanding

## Case Study – Understanding the data

- Descriptive statistics
  - Univariate statistics
  - Pairwise correlations
  - Histogram

$$f(a) + \sum_{k=1}^{n} \frac{1}{k!} \frac{d^k}{dt^k}\Big|_{t=0} f(u(t)) + \int_0^1 \frac{(1-t)^n}{n!} \frac{d^{n+1}}{dt^{n+1}} f(u(t))\, dt.$$

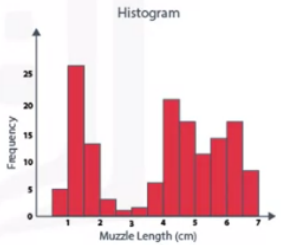
$F_{X,Y}(x, y)$ satisfies

$$F_{X,Y}(x, y) = F_X(x) F_Y(y),$$

or equivalently, their joint density $f_{X,Y}(x, y)$ satisfies

$$f_{X,Y}(x, y) = f_X(x) f_Y(y).$$

Histograms are a good way to understand how values or a variable are distributed, and what sorts of data preparation may be needed to make the variable more useful in a model.




BIG DATA UNIVERSITY

5

# Case Study - 5. Data Understanding

## Case study – Looking at data quality

- Data quality
  - Missing values
  - Invalid or misleading values

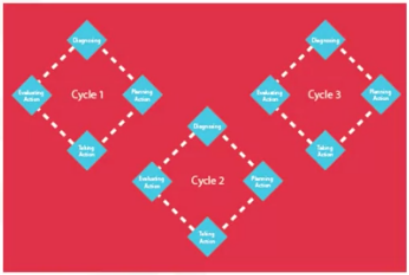

BIG DATA UNIVERSITY

6

# Case Study - 5. Data Understanding

## Case study – This is an iterative process

- Iterative data collection and understanding
  - Refined definition of "CHF admission"




BIG DATA UNIVERSITY

7

# 6. Data Preparation

- Most time consuming phase – 70% to 90% of overall project time
- Automating collection and preparation time can reduce to 50%.
- What are the ways in which data is prepared?
  - address missing or invalid values
  - remove duplicates
  - data to be properly formatted
- Transforming data
  - process of getting data into a state where it may be easier to work with.
- Feature Engineering
  - process of using domain knowledge f data to create features that make ML algo work.
  - Feature is a characteristic that might help solving a problem.

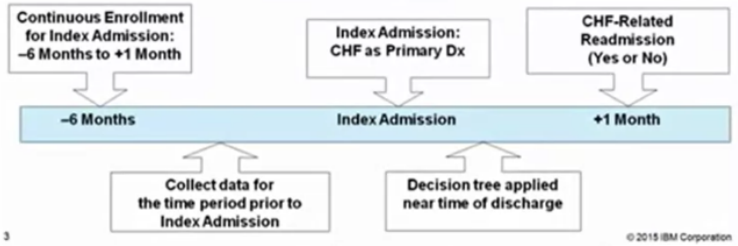
# Case Study - 6. Data Preparation

## Case Study – Defining readmission

# Case Study - 6. Data Preparation

## Case Study – Defining CHF admission

Define "CFF admission" and "CHF readmission"




BIG DATA UNIVERSITY

7

# Case Study - 6. Data Preparation

## Case Study – Aggregating records

Transactional records

- Claims: professional provider, facility, pharmaceutical
- Inpatient & outpatient records: diagnoses, procedures, prescriptions, etc.
- Possibly thousands per patient, depending on clinical history


BIG DATA UNIVERSITY

8

# Case Study - 6. Data Preparation

## Case Study – Aggregating to patient level

Aggregate to patient level

- Roll up to 1 record per patient
- Create new columns representing the transaction
  - Outpatients visits/ Inpatient episodes: frequency, recency, diagnoses/length of stay, procedures, prescriptions
  - Comorbidities with CHF


BIG DATA UNIVERSITY

9

# Case Study - 6. Data Preparation

## Case Study – More or less data needed?

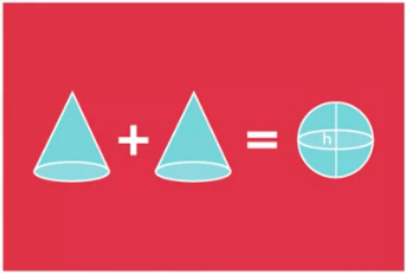
Literature review of important factors for CHF readmission

**MORE LIKELY TO BE READMITTED**

- Medicare / Medicaid insurance holders
- Comorbid conditions including:
  - Ischemic heart disease
  - Idiopathic Cardiomyopathy
  - Prior Cardiac surgery
  - Peripheral vascular disease
  - Diabetes mellitus
  - Anemia

**LESS LIKELY TO BE READMITTED**

- Patients treated at rural hospitals
- Patients discharged to skilled nursing facilities
- Patients receiving echocardiograms or cardiac catheterization




- Loop back to data collection stage and add additional data, if needed

BIG DATA UNIVERSITY

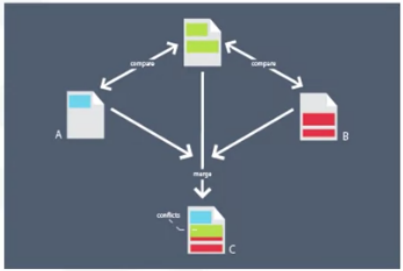
10

# Case Study - 6. Data Preparation

## Case Study – Completing the data set

Merge all data into one table

- One record per patient
- List of variables used in modeling
  - Target: CHF readmission with 30 days (Yes/No), following discharge from CHF hospitalization




BIG DATA UNIVERSITY

11

# Case Study - 6. Data Preparation

## Case Study – Creating new variables

Merge all data into one table

- One record per patient
- List of variables used in modeling

– Target

CHF readmission with 30 days (Yes/No), following discharge from CHF hospitalization

– Measures

| | | |
|---|---|---|
| Gender | Length of stay | CHF Diagnosis importance (primary, secondary, tertiary) |
| Age | Prior admissions | |
| Primary DRG | Line of business | |

– Diagnosis flags (Y/N)

| | | |
|---|---|---|
| CHF | Atrial fibrillation | Pneumonia |
| Diabetes | Renal failure | Hypertension |

BIG DATA UNIVERSITY

12

# Case Study - 6. Data Preparation

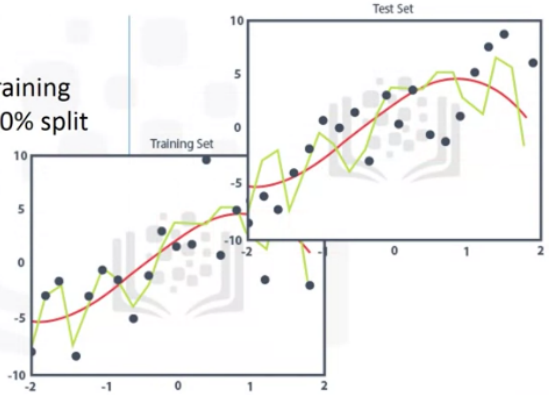
## Case Study – Using training sets

Cohort: 2,343 patients

Randomly divided into training and testing sets: 70% / 30% split

Training: 1,640 patients

Testing: 703 patients




BIG DATA UNIVERSITY

13

# 7. Modeling

- In what way can the data be visualized to get to the answer that is required?
- Data modeling focuses on developing models that are either descriptive or predictive.
- Modeling is based on the analyatic approach.
- Success of compilation, preparation and modeling depends on the understanding of problem and analytical approach being taken.
- Constant refinement, understanding and tweaking is essential.
- Relevance of the model.

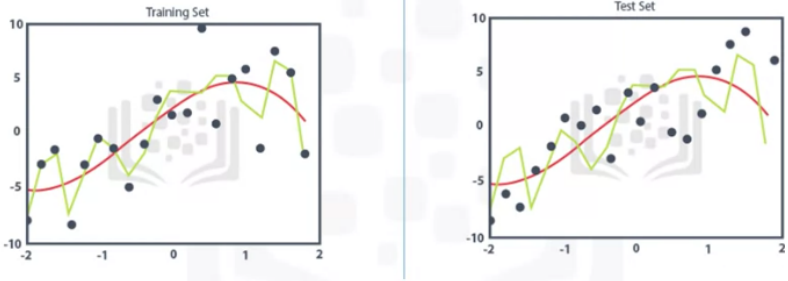
# 7. Modeling

## Data Modeling – Using Predictive or Descriptive?


BIG DATA UNIVERSITY

# 7. Modeling

## Data Modeling – Using training / test sets




Does the model need to be calibrated?

BIG DATA UNIVERSITY

# 7. Modeling

## Understanding the question


1. Understand the question at hand
2. Select an analytic approach or method to solve the problem
3. Obtain, understand, prepare, and model the data

BIG DATA UNIVERSITY

 7

# Case Study - 7. Modeling

## Case Study – Analyzing the 1st model

Initial decision tree classification model

- Low accuracy on "Yes" outcome

| Model | Relative Cost Y:N | Overall Accuracy (% correct Y & N) | Sensitivity (Y accuracy) | Specificity (N accuracy) |
|---|---|---|---|---|
| 1 | 1:1 | 85% | 45% | 97% |
| 2 | 9:1 | 49% | 97% | 35% |
| 3 | 4:1 | 81% | 68% | 85% |

BIG DATA UNIVERSITY

5

# 8. Evaluation

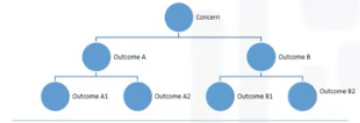
- Quality of the developed model is assessed.
- Before model gets deployed.
- Does the model really answer the inital question?
- Two phases
  - Diagnostic measure phase
    - ensure model is working as intended.
  - Statistical significance phase
    - ensure data is being properly handled and interpreted within the model.

# 8. Evaluation

## When and how to adjust the model?

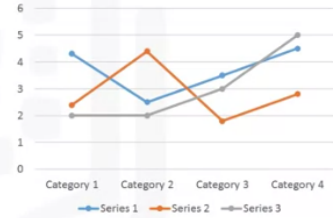
### Diagnostic measures

***Predictive Model***




***Descriptive Model***


***Statistical Significance***




BIG DATA UNIVERSITY

# Case Study - 8. Evaluation

## Case Study – Misclassification costs

Misclassification cost tuning

- Tune the relative misclassification costs
- Balance true-positive rate and false-positive rate for best model

| Model | Relative Cost Y:N | True Positive Rate (Sensitivity) | Specificity (accuracy on N) | False Positive Rate (1 – Specificity) |
|---|---|---|---|---|
| → 1 | 1:1 | 0.45 | 0.97 | 0.03 |
| 2 | 1.5:1 | 0.60 | 0.92 | 0.08 |
| 3 | 4:1 | 0.68 | 0.85 | 0.15 |
| 4 | 9:1 | 0.97 | 0.35 | 0.65 |

BIG DATA UNIVERSITY

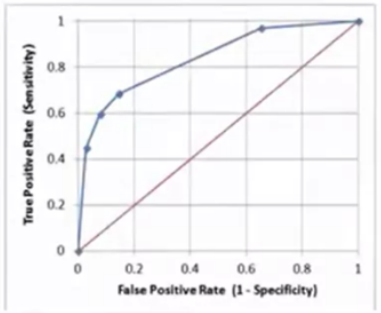
7

# Case Study - 8. Evaluation

## Case Study – Using the ROC curve

Diagnostic tool for classification model evaluation

- Classification model performance
- True-Positive Rate vs False-Positive Rate
- Optimal model at maximum separation




BIG DATA UNIVERSITY

11

# 9. Deployment

- Are the stakeholders familiar with the tool?
- Deploy the tested model.
- Roll out to a limited group of users or in a test environment.

# 9. Deployment

## Case Study – Understand the results

Assimilate knowledge for business

- Practical understanding of the meaning of model results
- Implications of model results for designing intervention actions

BIG DATA UNIVERSITY

6

# 9. Deployment

## Case Study – Gathering application requirements

Application requirements

- Automated, near-real-time risk assessments of CHF inpatients
- Easy to use
- Automated data preparation and scoring
- Up-to-date risk assessment to help clinicians target high-risk patients

BIG DATA UNIVERSITY

7

# 9. Deployment

## Case Study – Additional requirements?

Additional requirements
- Training for clinical staff
- Tracking / monitoring processes

BIG DATA UNIVERSITY

8

# 10. Feedback

- Feedback from users to refine the model
- Assess the model for performance and impact.

# 10. Feedback

## Case Study – Assessing model performance

Define review process

- To measure results of applying the risk model to the CHF patient population
- Track patients who received intervention
  - Actual readmission outcomes
- Measure effectiveness of intervention
  - Compare readmission rates before & after model implementation


BIG DATA UNIVERSITY

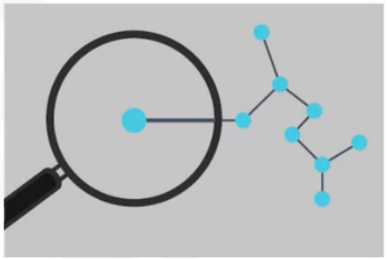
7

# 10. Feedback

## Case Study – Refinement

Refine model

- Initial review after the first year of implementation
- Based on feedback data and knowledge gained
- Participation in intervention program
- Possibly incorporate detailed pharmaceutical data originally deferred
- Other possible refinements as yet unknown



BIG DATA UNIVERSITY

8

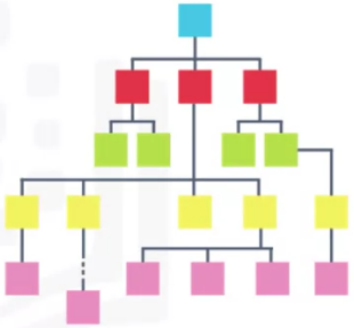
# 10. Feedback

## Case Study – Redeployment

Review and refine intervention actions

Redeploy

- Continue modeling, deployment, feedback, and refinement throughout the life of the intervention program



BIG DATA UNIVERSITY

9

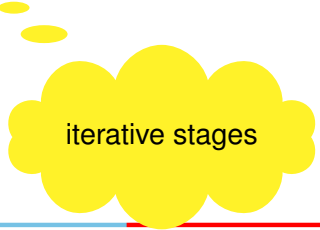
# Data Science Process

- Learn the importance of
  - Understanding the question
  - Picking the most effective analytic approach
- Learn to work with data
  - determine the data requirements
  - collect the appropriate data
  - understand the data
  - prepare the data for modeling
- Learn how to
  - evalaute and deploy the model
  - getting feedback on it
  - use the feedback constructively so as to impove the model

All stages of the methodology are iterative.

# Data Science Process

- Think like a data scientist
  - Forming a concrete business or research problem
  - Collecting and analyzing data
  - Building a model
  - Understanding the feedback after model deployment



# THANK YOU