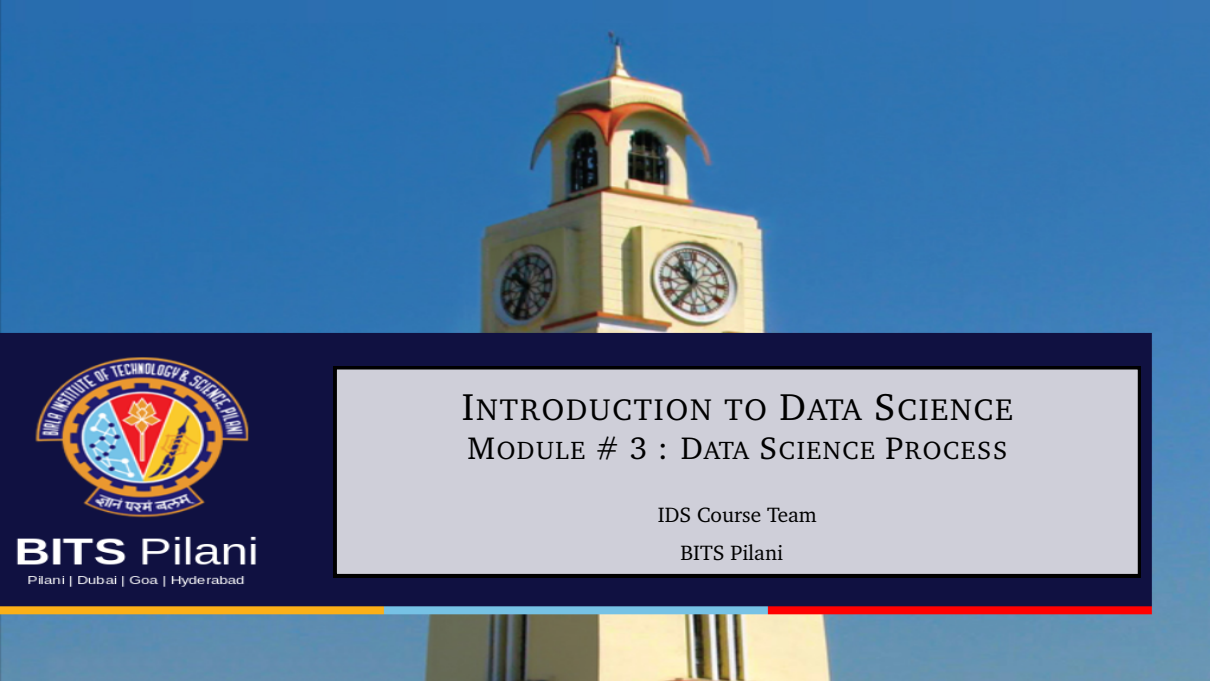


BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Data Science
Module # 3 : Data Science Process

IDS Course Team

BITS Pilani

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

Table of Contents

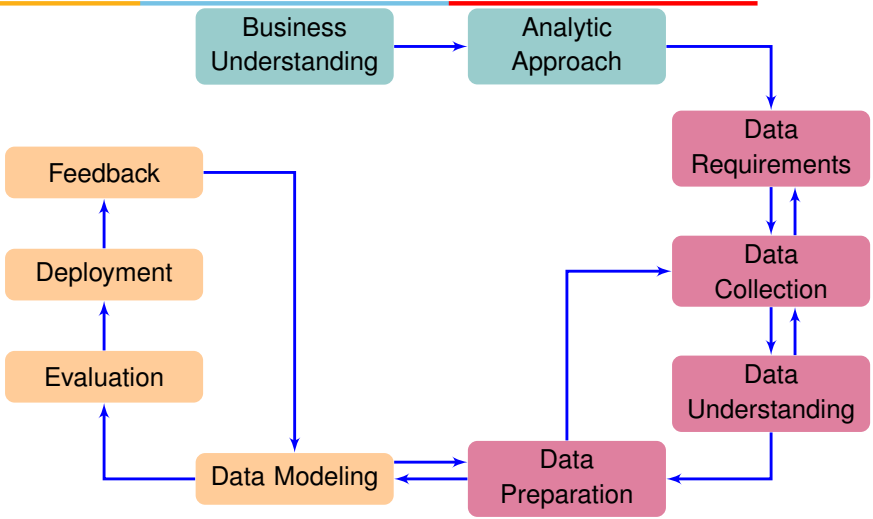
Data Science Process

- 10 Questions the process aims to answer
  - 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

Source: CognitiveClass

Data Science Process



Data Science Process

- From Problem to Approach
  - Business Understanding
  - Analytic Approach
- From Requirements to Collection
  - Data Requirements
  - Data Collection
- From Understanding to Preparation
  - Data Understanding
  - Data Preparation
- From Modeling to Evaluation
  - Modeling
  - Evaluation
- From Deployment to Feedback
  - Deployment
  - Feedback

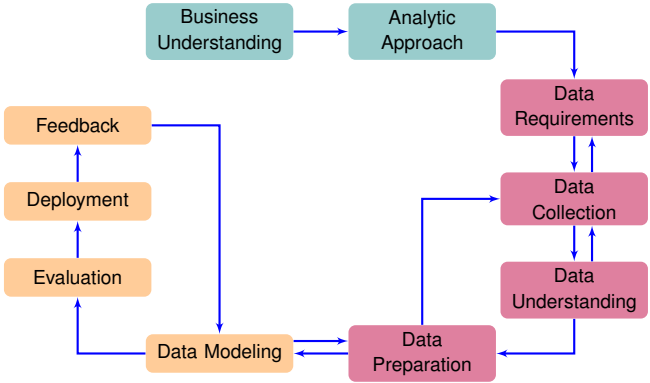


Table of Contents

Hospital Readmissions

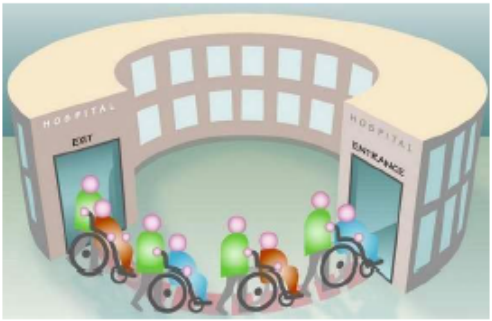


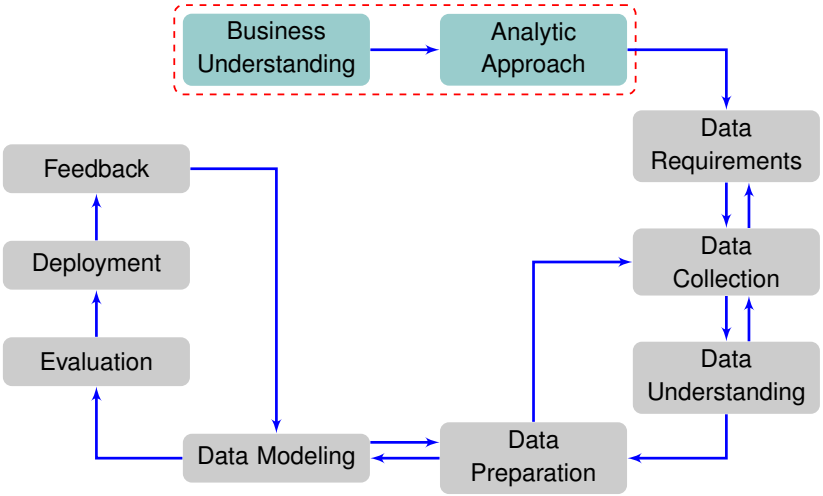
Image Source:
https://medium.com/nwamaka-imasogie/predicting-hospital-readmission-using-nlp-5f0fe6f1a705

Hospital Readmissions - Scenario

- There is a limited budget for providing healthcare to the public.
- Hospital readmissions for re-occurring problems are considered as a sign of failure in the healthcare system.
- There is a dire need to properly address the patient condition prior to the initial patient discharge.
- The core question is:
  - What is the best way to allocate these funds to maximize their use in providing quality care?
- A successful data science program will deliver:
  - better patient care by giving physicians new tools to incorporate timely, data-driven information into patient care decisions.

Source: CognitiveClass

From Problem to Approach



Data Problem to Analytic Approach

- The need to understand and prioritize the business goal.
- The way stakeholder support influences a project.
- The importance of selecting the right model.
- When to use a predictive, descriptive, or classification model.

Source: CognitiveClass

1. Business Understanding (Concept)

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

Source: CognitiveClass

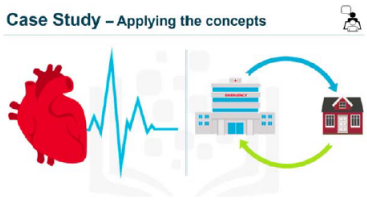
Case Study - 1. Business Understanding

- Case study asks the following question:
  - What is the best way to allocate the limited healthcare budget to maximize its use in providing quality care?
- Goals and Objectives
  - Define the GOALS.
    - Provide quality care without increasing cost.
  - Define the OBJECTIVES.
    - Review the process to identify inefficiencies.

Source: CognitiveClass

Case Study - 1. Business Understanding

- Examining hospital readmissions
  - It was found that approximately 30% of individuals who finish rehab treatment would be readmitted to a rehab center within one year.
  - 50% would be readmitted within five years.
  - After reviewing some records, it was found that patients with heart failure were high on the list of readmission.



Source: CognitiveClass

Case Study - 1. Business Understanding

- It was found that a decision tree model can be applied to investigate this scenario to determine the reason for this phenomenon.
- It is important to gain business insight for the analytics team to implement the data science project.
  - Data scientists proposed and organized an on-site workshop.
- The business sponsors involvement throughout the project was critical because the sponsor had
  - Set the overall direction
  - Remained committed and advised
  - when required, got the necessary support

Source: CognitiveClass

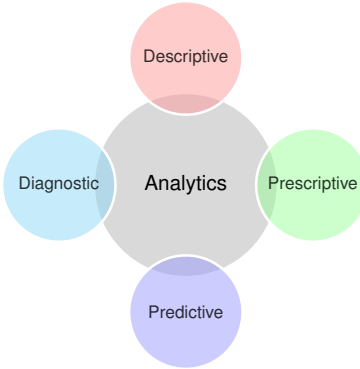
Case Study - 1. Business Understanding

- Finally, four business requirements were identified for whatever model would be built
  - Case study question
    - What is the best way to allocate the limited healthcare budget to maximize its use in providing quality care?
  - Business requirements
    - To predict the risk of readmission.
    - To predict readmission outcomes for those patients with Congestive Heart Failure.
    - To understand the combination of events that led to the predicted outcome.
    - To apply an easy-to-understand process to new patients, regarding their readmission risk.

Source: CognitiveClass

2. Analytic Approach (Concept)

- Available data
  - Patient data, Readmissions data, CFH data, etc
- 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?



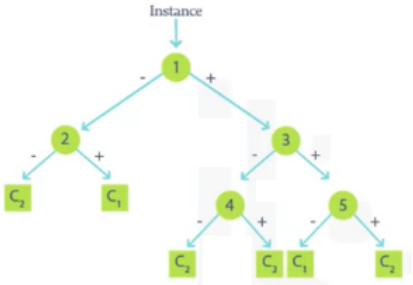
2. Analytic Approach (Concept)

- The analytic approach can be selected once a clear understanding of the question is established:
  - If the question is to determine probabilities of an action.
    - Predictive model can be used
  - If the question is to show relationships
    - Use a Descriptive model
  - If the question requires a yes / no answer
    - Classification approach to predicting a response is appropriate.

Source: CognitiveClass

Case Study - 2. Analytic Approach

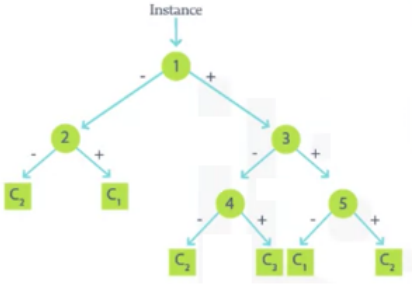
- A decision tree classification model was used to identify the combination of conditions leading to each patient's outcome.
- Examining the variables in each of the nodes along each path to a leaf, led to a respective threshold value to split the tree. Eg: $Age >= 60$
- A decision tree classifier provides both the predicted outcome, as well as the likelihood of that outcome, based on the proportion at the dominant outcome, yes or no, in each group.



Source: CognitiveClass

Case Study - 2. Analytic Approach

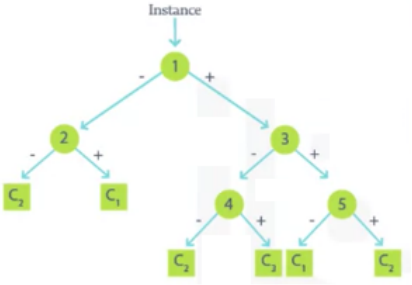
- The analysts can obtain the readmission risk, or the likelihood of a yes for each patient.
- If the dominant outcome is yes, then the risk is simply the proportion of yes patients in the leaf.
- If it is no, then the risk is 1 minus the proportion of no patients in the leaf.
- For non-data scientists, a decision tree classification model is easy to understand and apply, to score new patients for their risk of readmission.



Source: CognitiveClass

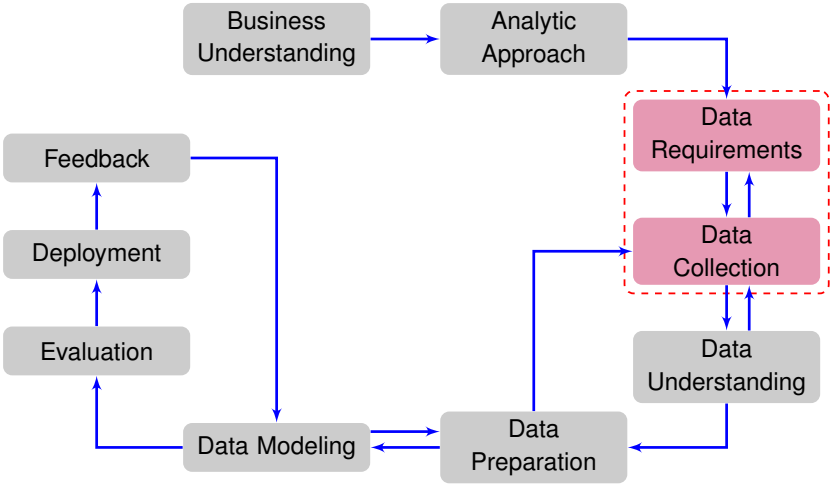
Case Study - 2. Analytic Approach

- Clinicians can readily see what conditions are causing a patient to be scored as high-risk.
- Multiple models can be built and applied at various points during hospital stay.
- This gives a moving picture of the patient's risk and how it is evolving with the various treatments being applied.
- For these reasons, the decision tree approach was chosen for building the Congestive Heart Failure (CHF) readmission model.



Source: CognitiveClass

From Data Requirements to Data Collection



From Data Requirements to Data Collection

- The significance of defining the data requirements for the model.
- Why the content, format, and representation of data matter
- The importance of identifying the correct sources for data collection
- How to handle unavailable and redundant data.
- To anticipate the needs of future stages in the process.

Source: CognitiveClass

3. Data Requirements (Concept)

- If our goal is to make a ”Biryani” but we don’t have the right ingredients, then the success of making a good Biryani will be compromised.
- If the ”recipe” is the problem to be solved, then the data are the ingredients.
- The data scientist must ask the following questions:
  - What are the data requirements?
  - How to obtain or collect them?
  - How to understand and use them?
  - How to prepare the data to meet the desired outcome?
- Based on the understanding of the problem and the analytic approach chosen, it is important to define the data requirements.

Source: CognitiveClass

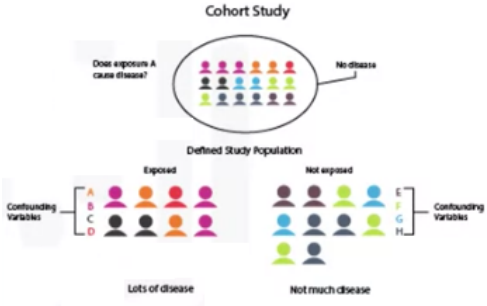
Case Study - 3. Data Requirements

- The analytic approach is decision tree classification, so data requirements should be defined accordingly.
- This involves:
  - Identify data content
  - Identify data formats
  - Identify data sources needed for the initial data collection.

Source: CognitiveClass

Case Study - 3. Data Requirements

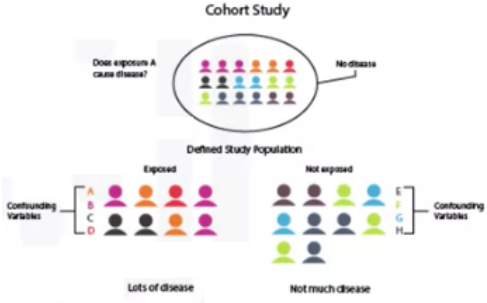
- Data requirements for the case study included selecting a suitable list of patients from the health insurance providers member base.
- In order to put together patient clinical histories, three criteria were identified for selecting the patient cohort.
  1. A patient must be admitted as an in-patient within health insurance provider's service area.
  2. Patient's primary diagnosis should be CHF for one full year.
  3. Prior to the primary admission for CHF, a patient must have had at least 6 months of continuous enrollment.



Source: CognitiveClass

Case Study - 3. Data Requirements

- Disqualifying conditions (outliers)
  - CHF patient who have been diagnosed with other serious conditions are excluded because this may result in above-average rates of re-entry and may therefore distort results.



Source: CognitiveClass

Case Study - 3. Data Requirements

Defining the data

- The content and format suitable for decision tree classifier needs to be defined.
- Format
  - Transactional format
  - This model requires, one record per patient.
  - Columns of the record represent dependent and predictor variables.
- Content
  - To model the readmission outcome, data should represent all aspects of the patient's clinical history.
  - This includes:
    - ★ Authorizations
    - ★ Primary, secondary and tertiary diagnoses,
    - ★ procedures, prescriptions and other services provided during hospitalization or visits by patients / doctors.

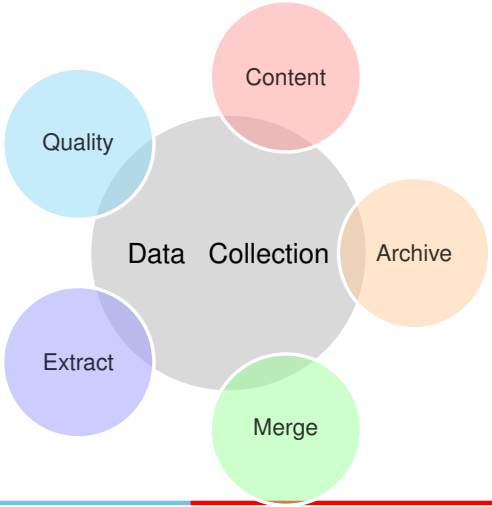
Case Study - 3. Data Requirements

- A given patient can have thousands of records that represent all their attributes.
- The data analytics specialists collected the transaction records from patient records and created a set of new variables to represent that information.
- It was a task for the data preparation phase, so it is important to anticipate the next phases.

Source: CognitiveClass

4. Data Collection (Concept)

- What occurs during data collection?
  - Decide 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?



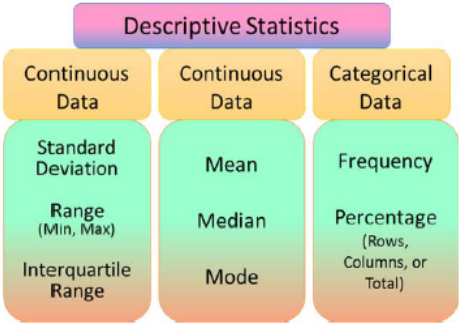
4. Data Collection (Concept)

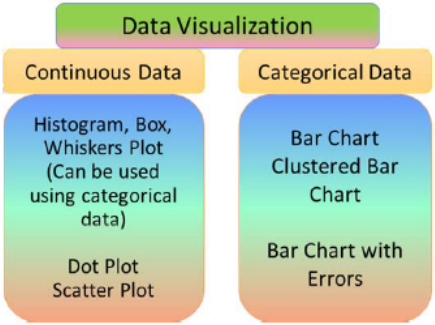
- Once data collection is completed, the Data Scientist performs an assessment to make sure he has all the required data.
- As with the purchase of ingredients for making a meal, some ingredients may be out of season and more difficult to obtain or cost more than originally planned.
- At this stage, the data requirements are reviewed and a decision is made as to whether more or less data is required for the collection.
- The gaps in the data are identified and plans for filling or replacement must be made.
- Once this step is complete, essentially, the ingredients are now ready for washing and cutting.

Source: CognitiveClass

4. Data Collection (Concept)

- The collected data is explored using descriptive statistics and visualization to assess its content and quality.





Case Study - 4. Data Collection

- This case study required data about:
  - Demographics, clinical and coverage information of patients, provider information, claims records, as well as pharmaceutical and other information related to all the diagnoses of the CHF patients.
- Available data sources
  - Corporate data warehouse
    - ★ Single source of medical, claims, eligibility, provider, and member information.
  - In-patient record system
  - Claim patient system
  - Disease management program information

Source: CognitiveClass

Case Study - 4. Data Collection

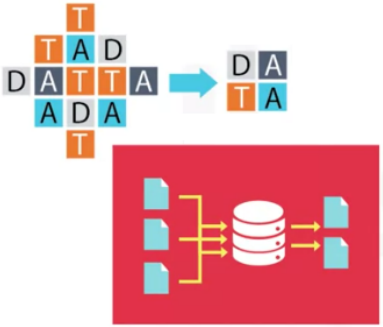
- This case study also required other data, but not available.
  - Pharmaceutical records
  - Information on drugs

- This data source was not yet integrated with the rest of the data sources.
- In such situations,
  - It is okay to postpone decisions about unavailable data and to try to capture them later.
  - This can happen even after obtaining intermediate results from predictive modeling.
  - If the results indicate that drug information may be important for a good model, you will spend time trying to get it.
- However, it turned out that they could build a reasonably good model without this information about drugs.

Source: CognitiveClass

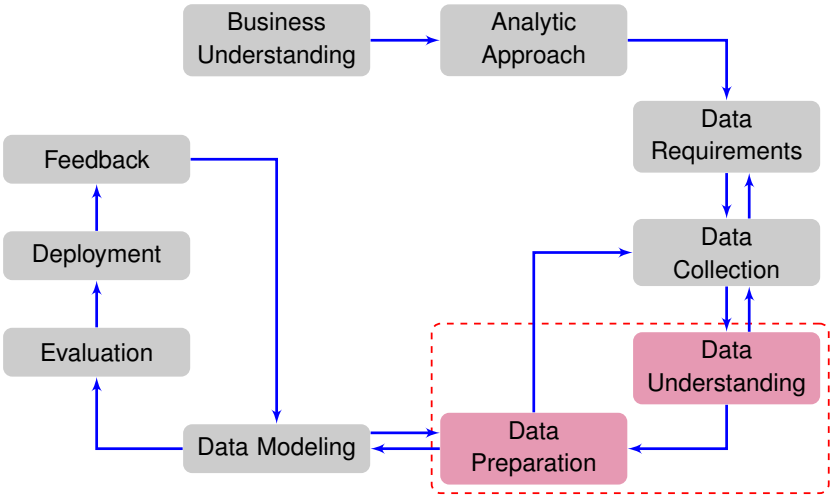
Case Study - 4. Data Collection

- Data Pre-processing and Merging Data
  - Database administrators and programmers often work together to extract data from different sources and then combine them.
  - Redundant data can be deleted and made available to the next level of methodology – the ”Data Understanding” phase.
  - At this stage, scientists and analysts can discuss ways to better manage their data by automating certain database processes to facilitate data collection
- Next, we move on to understanding the data



Source: CognitiveClass

From Data Understanding to Data Preparation



From Data Understanding to Data Preparation

- The importance of descriptive statistics.
- How to manage missing, invalid, or misleading data?
- The need to clean data and sometimes transform data.
- The consequences of bad data for the model.
- Data understanding is iterative.
  - We learn more about data, the more we study it.

Source: CognitiveClass

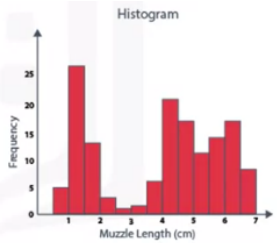
5. Data Understanding (Concepts)

- This section of the methodology answers the question.
  - 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
- From the data collected, we should understand the variables and their characteristics using Exploratory Data Analysis and Descriptive Statistics.
- Sometimes we may have to perform pre-processing operations on the data.

Source: CognitiveClass

Case Study - 5. Data Understanding

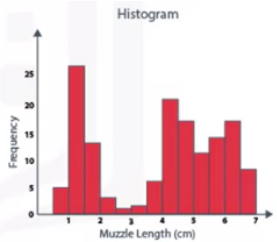
- First, Univariate Statistics
  - Basic statistics included univariate statistics for each variable, such as:
    - mean, median, minimum, maximum, and standard deviation
- Second, Pairwise Correlations
  - Pairwise correlations were used to determine the degree of correlation between the variables.
  - Variables that are highly correlated means they are essentially redundant.
    - This makes only one variable relevant for the modeling.



Source: CognitiveClass

Case Study - 5. Data Understanding

- Third, Histograms
  - Third, the histograms of the variables were examined to understand their distributions.
  - Histograms are a good way to understand how values or variables are distributed.
  - They help to know what kind of data preparation may be needed to make the variable more useful in a model.
  - For example:
    - ⋆ If a categorical variable contains too many different values to be meaningful in a model, the histogram can help decide how to consolidate those values.



Source: CognitiveClass

Case Study - 5. Data Understanding

- Looking at data quality
  - Univariate, statistics and histograms are also used to assess the quality of the data.
  - On the basis of the data provided, some values can be recoded or deleted if necessary.
    - E.g., if a particular variable has a lot of missing values, we may drop the variable from the model.
  - Sometimes a missing value means ”no” or ”0” (zero), or sometimes simply ”we do not know”.
-
  - A variable contains invalid or misleading values.
    - E.g., A numeric variable called ”age” containing 0 to 100 and 999, where ”triple-9” actually means ”missing”, will be treated as a valid value unless we have corrected it.

Source: CognitiveClass

Case Study - 5. Data Understanding

- Data understanding is an iterative process.
  - Originally, the meaning of CHF admission was decided on the basis of a primary diagnosis of CHF.
  - However, preliminary data analysis and clinical experience revealed that CHF admissions were also based on other diagnosis.
    - ★ The initial definition did not cover all cases of CHF admissions.
  - They added secondary and tertiary diagnoses, and created a more complete definition of CHF admission.
  - This is one example of the iterative processes in the methodology .
  - The more we work with the problem and the data, the more we learn and the more the model can be adjusted, which ultimately leads to a better resolution of the problem.

Source: CognitiveClass

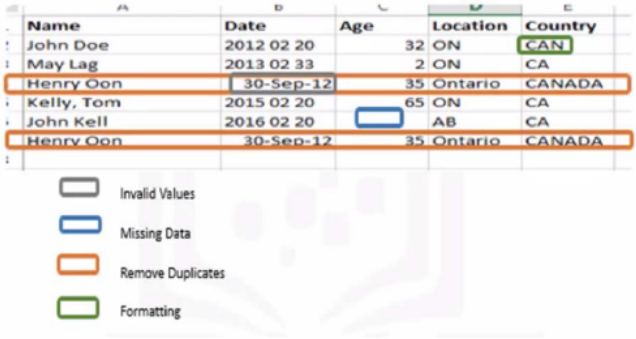
6. Data Preparation (Concept)

- In a way, data preparation is like removing dirt and washing vegetables.
- Compared to data collection and understanding, data preparation is the most time consuming phase – 70% to 90% of overall project time.
- Automating collection and preparation time can reduce to 50%.
- The data preparation phase of the methodology answers the question:
  - What are the ways in which data is prepared?
    - Address missing or invalid values
    - Remove duplicates
    - Format data properly
- Transforming data
  - Process of getting data into a state where it may be easier to work with.
- Feature Engineering

Source: CognitiveClass

Case Study - 6. Data Preparation

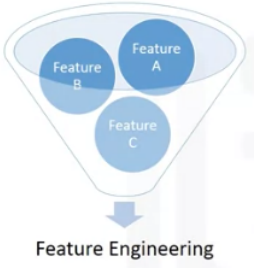
Examples of data cleansing



Source: CognitiveClass

6. Data Preparation (Concept)

- Feature Engineering
  - Process of using domain knowledge of data to create features that make ML algorithms work.
  - Feature is a characteristic that might help solving a problem.
  - Feature engineering is also part of the data preparation.
  - Use domain knowledge on data to create features that work with machine learning algorithms.
  - A feature is a property that can be useful for solving a problem.
  - The functions in the data are important for the predictive models and influence the desired results.



Source: CognitiveClass

Case Study - 6. Data Preparation

- Defining Congestive Heart Failure
  - In the case study, first step in the data preparation stage was to actually define what CHF means.
  - First, the set of diagnosis-related group codes needed to be identified, as CHF implies certain kinds of fluid buildup.
  - Data scientists also needed to consider that CHF is only one type of heart failure.
  - Clinical guidance was needed to get the right codes for CHF.

Source: CognitiveClass

Case Study - 6. Data Preparation

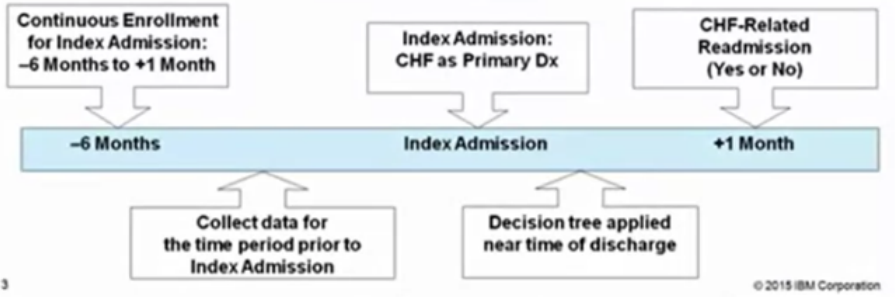
- Defining re-admission criteria for Congestive Heart Failure
  - Next step involved defining the criteria for CHF readmissions.
  - The timing of events needed to be evaluated in order to define whether a particular CHF admission was an initial event (called as index admission), or a CHF-related re-admission.
  - Based on clinical expertise, a time period of 30 days was set as the window for readmission relevant for CHF patients, following the discharge from the initial admission.

Source: CognitiveClass

Case Study - 6. Data Preparation

Case Study – Defining CHF admission

Define "CFF admission" and "CHF readmission"



Case Study - 6. Data Preparation

- Aggregating transactional records
  - Next, the records that were in transactional format were aggregated.
  - Meaning that the data included multiple records for each patient.
  - Transactional records included claims submitted for physician, laboratory, hospital, and clinical services.
  - Also included were records describing all the diagnoses, procedures, prescriptions, and other information about in-patients and out-patients.

Source: CognitiveClass

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

Source: CognitiveClass

Case Study - 6. Data Preparation

- Aggregating data to patient level
  - A given patient could have hundreds or even thousands of records, depending on their clinical history.
  - All the transactional records were aggregated to the patient level, yielding a single record for each patient.
    - This is required for the decision-tree classification method used for modeling.
  - Many new columns were created representing the information in the transactions.
    - E.g: Frequency and most recent visits to doctors, clinics and hospitals with diagnoses, procedures, prescriptions, and so forth.
  - Co-morbidities with CHF were also considered, such as:
    - Diabetes, hypertension, and many other diseases and chronic conditions that could impact the risk of re-admission for CHF.

Source: CognitiveClass

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

Source: CognitiveClass

Case Study - 6. Data Preparation

- Do we need more data or less data?
  - A literature review on CHF was also undertaken to see whether any important data elements were overlooked.
    - Such as co-morbidities that had not yet been accounted for.
  - The literature review involved looping back to the data collection stage to add a few more indicators for conditions and procedures.

Source: CognitiveClass

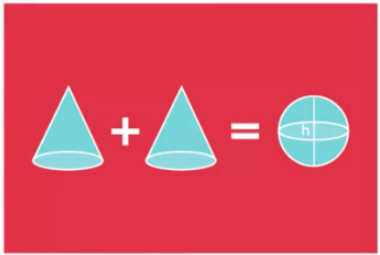
Case Study - 6. Data Preparation

Case Study – More or less data needed?

Literature review of important factors for CHF readmission

MORE LIKELY TO BE READMITTED	LESS 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	• 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



Source: CognitiveClass

Case Study - 6. Data Preparation

- Creating new variables
  - Aggregating the transactional data at the patient level, involved.
    - merging it with the other patient data, including their demographic information, such as age, gender, type of insurance, and so forth.
  - The result was the creation of one table containing a single record per patient.
  - Columns represent the attributes about the patient in his or her clinical history.
  - These columns would be used as variables in the predictive modeling.

Source: CognitiveClass

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
Age	Prior admissions	(primary, secondary, tertiary)
Primary DRG	Line of business	

– Diagnosis flags (Y/N)

CHF	Atrial fibrillation	Pneumonia
Diabetes	Renal failure	Hypertension

Source: CognitiveClass

Case Study - 6. Data Preparation

Completing the data set

- Here is a list of the variables that were ultimately used in building the model
  - Measures
    - Gender, Age, Primary Diagnosis Related Group (DRG), Length of Stay, CHF Diagnosis Importance (primary, secondary, tertiary), Prior admissions, Line of business.
  - Diagnosis Flags (Y/N)
    - CHF, Atrial fibrillation, Pneumonia, Diabetes, Renal failure, Hypertension.
- Dependent Variable
  - CHF readmission within 30 days following discharge from CHF hospitalization (Yes/No).

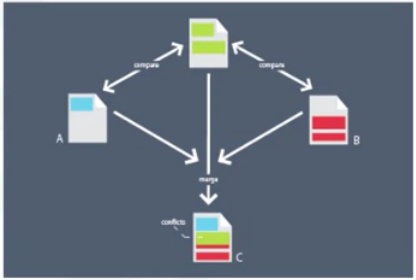
Source: CognitiveClass

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



Source: CognitiveClass

Case Study - 6. Data Preparation

- Creating training and testing datasets
  - The data preparation stage resulted in a cohort of 2,343 patients.
  - These patients met all of the criteria for this case study.
  - The data (patient records) were then split into training and testing sets for building and validating the model, respectively.

Source: CognitiveClass

Case Study - 6. Data Preparation

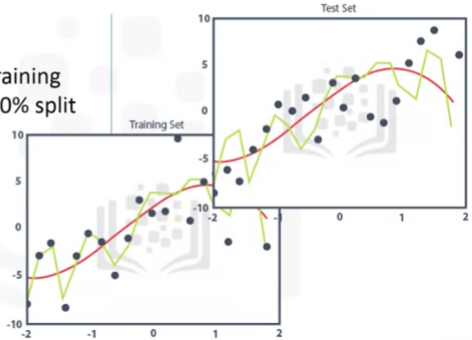
Case Study – Using training sets

Cohort: 2,343 patients

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

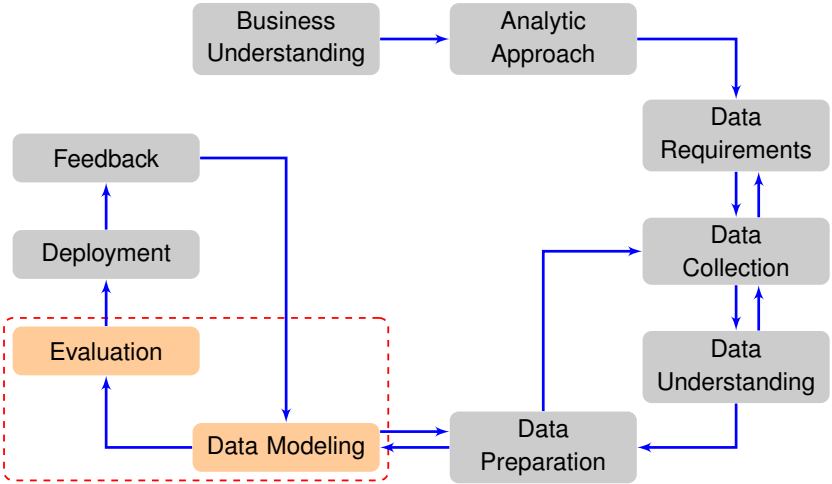
Training: 1,640 patients

Testing: 703 patients



Source: CognitiveClass

From Data Modeling to Evaluation



From Data Modeling to Evaluation

- The difference between descriptive and predictive models.
- The role of training sets and test sets.
- The importance of asking if the question has been answered.
- Why diagnostic measures tools are needed.
- The purpose of statistical significance tests.
- That modeling and evaluation are iterative processes.

Source: CognitiveClass

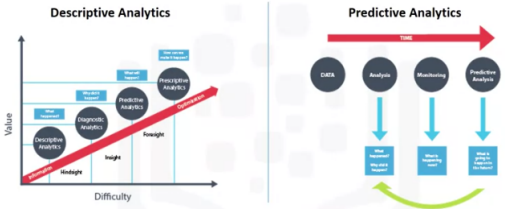
7. Data Modeling (Concept)

- In what way can the data be visualized to get to the answer that is required?
- Modeling is based on the analyatic approach.
- Data modeling focuses on developing models that are either descriptive or predictive.
  - Descriptive Models
    - What happened?
    - Use statistics.
  - Predictive Models
    - What wil happen?
    - Use machine learning.
    - Try to generate yes/no type outcomes.
    - A training set is used for developing the predictive model.
  - Training set
    - contains historical data in which the outcomes are already known.
    - acts like a gauge to determine if the model needs to be calibrated.

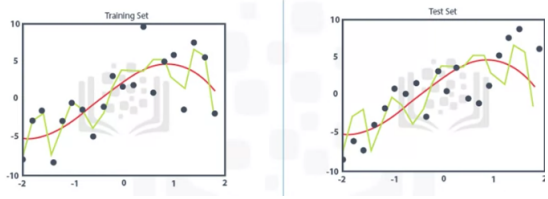
Source: CognitiveClass

7. Data Modeling (Concept)

Data Modeling – Using Predictive or Descriptive?



Data Modeling – Using training / test sets



Source: CognitiveClass

7. Data Modeling (Concept)

- The data scientist will try different algorithms to ensure that the variables in play are actually required.
- Success of compilation, preparation and modeling depends on the understanding of problem and analytical approach being taken.
- Like the quality of ingredients in cooking, the quality of data sets the stage for the outcome.
  - If data quality is bad, the outcome will be bad.
- Constant refinement, adjustment, and tweaking within each step are essential to ensure a solid outcome.
- The end goal is to build a model that can answer the original question.
  - Model evaluation, deployment, and feedback loops ensure that the model is relevant and the question is really answered.

Source: CognitiveClass

Case Study - 7. Data Modeling

- In this first model, the default is 1-to-1 is used.
- The overall accuracy in classifying the yes and no outcomes was 85%.
- This sounds good, but it represents only 45% of the ”yes”.
  - Meaning, when it’s actually YES, model predicted YES only 45% of the time.
- The question is:
  - How could the accuracy of the model be improved in predicting the yes outcome?

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%

Source: CognitiveClass

Case Study - 7. Data Modeling

- There are many aspects to model building – one of those is parameter tuning to improve the model.
- With a prepared training set, the first decision tree classification model for CHF readmission can be built.
- We are looking for patients with high-risk readmission, so the outcome of interest will be CHF readmission equals ”yes”.
- For decision tree classification, the best parameter to adjust is the relative cost of misclassified yes and no outcomes.

Source: CognitiveClass

Case Study - 7. Data Modeling

- Type I Error or False positive
  - When a true, non-readmission is misclassified, and action is taken to reduce that patient's risk, the cost of that error is the wasted intervention.
- Type II Error or False negative
  - When a true readmission is misclassified, and no action is taken to reduce that risk.
  - The cost of this error is the readmission and all its attended costs, plus the trauma to the patient.

		Actual Condition	
		Positive	Negative
Predicted Condition	Positive	True Positive (Power)	False Positive (Type I Error)
	Negative	False Negative (Type II Error)	True Negative

- The costs of the two different kinds of misclassification errors can be quite different.
  - Adjust the relative weights of misclassifying the yes and no outcomes.
- For decision tree classification, the best parameter to adjust is the relative cost of misclassified yes and no outcomes.

Case Study - 7. Data Modeling

- For the second model, the relative cost was set at 9-to-1.
  - Ratio of cost of false positive to false negative.
  - This is a very high ratio, but gives more insight to the model's behavior.
- This time the model correctly classified 97% of the YES, but at the expense of a very low accuracy on the NO, with an overall accuracy of only 49%.
- This was clearly not a good model.
- The problem with this outcome is the large number of false-positives.
  - A true, non-readmission is misclassified as re-admission.
  - This would recommend unnecessary and costly intervention for patients, who would not have been re-admitted anyway.

Case Study – Analyzing the 2nd model

Second model
- High accuracy on "Yes" but poor on "No"

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%

Source: CognitiveClass

Case Study - 7. Data Modeling

- Try again to find a better balance between the yes and no accuracies.
- For the third model, the relative cost was set at 4-to-1.
- This time, the overall accuracy was 81%.
- Yes accuracy was 68%. This is called sensitivity.
- No accuracy was 85%. This is called specificity.
- This is the optimum balance that can be obtained with a rather small training set.
  - By adjusting the relative cost of misclassified yes and no outcomes parameter.
- In medical diagnosis
  - Test sensitivity is the ability of a test to correctly identify those with the disease (true positive rate).
  - Test specificity is the ability of the test to correctly identify those without the disease (true negative rate).

Case Study – Analyzing the 3rd model

Third model
- Better balance on "Yes" and "No" accuracy

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%

Confusion Matrix

- Confusion matrix is a table that is often used to evaluate the performance of a classification model (or "classifier").
- It works on a set of test data for which the true values are known.
- There are two possible predicted classes: "YES" and "NO".
- If we were predicting the presence of a disease, for example, "yes" would mean they have the disease, and "no" would mean they don't have the disease.
-
  - The classifier made a total of 165 predictions.
  - 165 patients were being tested for the presence of that disease.
  - Out of those 165 cases, the classifier predicted "yes" 110 times, and "no" 55 times.
  - In reality, 105 patients in the sample have the disease, and 60 patients do not.

N = 165	Predicted: No	Predicted Yes
Actual: No:	50	10
Actual: Yes:	5	100

Confusion Matrix

- True positives (TP) / Sensitivity:
  - The model predicted yes, and the patients have the disease.
- True negative (TN) / Specificity:
  - The model predicted no, and the patients don't have the disease.
- False positives (FP) / Type I error:
  - The model predicted YES, but the patients don't actually have the disease.
- False negatives (FN) / Type II error:
  - The model predicted NO, but the patients actually have the disease.

N = 165	Predicted: No	Predicted Yes
Actual: No:	TN = 50	FP =10
Actual: Yes:	FN = 5	TP = 100

Confusion Matrix

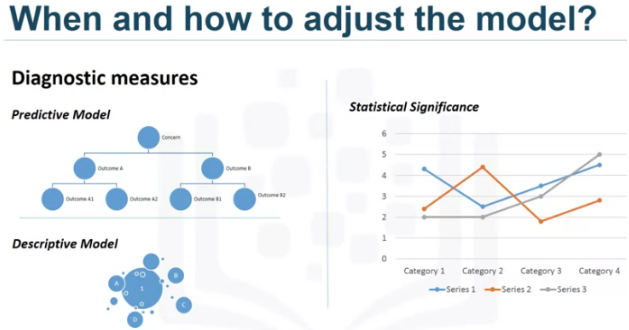
Term	Description	Calculation
Accuracy	Overall, how often is the classifier correct?	(TP+TN)/total = (100+50)/165 = 0.91
Misclassification Rate Error Rate	Overall, how often is it wrong? Equivalent to 1 minus Accuracy	(FP+FN)/total = (10+5)/165 = 0.09
True Positive Rate (Sensitivity or Recall)	When it's actually YES, how often does it predict YES?	TP/actual YES = 100/105 = 0.95
True Negative Rate (Specificity)	When it's actually NO, how often does it predict NO? Equivalent to 1 minus False Positive Rate	TN/actual NO = 50/60 = 0.83

Confusion Matrix

Term	Description	Calculation
False Positive Rate (Type I Error)	When it's actually NO, how often does it predict YES?	FP/actual NO = 10/60 = 0.17
True Negative Rate (Specificity)	When it's actually NO, how often does it predict NO? Equivalent to 1 minus False Positive Rate	TN/actual NO = 50/60 = 0.83
Precision	When it predicts YES, how often is it correct?	TP/predicted YES = 100/110 = 0.91
Prevalence	How often does the YES condition actually occur in our sample?	Actual YES/total = 105/165 = 0.64

8. Evaluation (Concept)

- Quality of the developed model is assessed.
- Before model gets deployed, evaluate whether the model really answers the initial question.



8. Evaluation (Concept)

Two phases

- Diagnostic measure phase
  - Ensures that the model works as intended.
  - If the model is a predictive model
    - ⋆ A decision tree can be used to assess whether the response provided by the model matches the original design.
    - ⋆ This allows areas to be displayed where adjustments are required.
  - If the model is a descriptive model that evaluates relationships
    - ⋆ A set of tests with known results can be applied and the model refined as necessary.
- Statistical significance phase
  - applied to the model to ensure data is being properly handled and interpreted within the model.
  - This is to avoid a second unnecessary assumption when the answer is revealed

Source: CognitiveClass

Case Study - 8. Evaluation

- One way is to find the optimal model through a diagnostic measure based on tuning one of the parameters in model building.
- Specifically we'll see how to tune the relative cost of misclassifying yes and no outcomes.
- Four models were built with four different relative misclassification costs.
- Each value of this model-building parameter increases the true positive rate of the accuracy in predicting yes, at the expense of lower accuracy in predicting no, that is, an increasing false-positive rate.

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

Source: CognitiveClass

Case Study - 8. Evaluation

- Which model is best based on tuning this parameter?
- Risk-reducing intervention – two scenarios
  - Cannot be applied to all CHF patients because many of them would not have been readmitted anyway. This will be cost effective.
  - The intervention itself would not be as effective in improving patient care if not enough high-risk CHF patients targeted.
- How do we determine which model was optimal?
  - This can be done with the help of an ROC curve (receiver operating characteristic curve).
- ROC curve is a graph showing the performance of a classification model at all classification thresholds.
- ROC curve plots two parameters:
  - True Positive Rate
  - False Positive Rate

Receiver Operator Characteristic (ROC) Curve

- ROC curves are used to show the connection/trade-off between clinical sensitivity and specificity for every possible cut-off (threshold) for a test or a combination of tests.
- The area under an ROC curve is a measure of the usefulness of a test in general.
  - A greater area means a more useful test.
- ROC curves are used in clinical biochemistry to choose the most appropriate cut-off for a test.
- The best cut-off has the **highest** true positive rate together with the **lowest** false positive rate.
- ROC curves were first employed in the study of discriminator systems for the detection of radio signals in the presence of noise in the 1940s, following the attack on Pearl Harbor.
- The initial research was motivated by the desire to determine how the US RADAR "receiver operators" had missed the Japanese aircraft.

Receiver Operator Characteristic (ROC) Curve

- An ROC curve plots TPR vs. FPR at different classification thresholds.
- Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives.
- The optimal model is the one giving the maximum separation between the blue ROC curve relative to the red base line.
- This curve quantifies how well a binary classification model performs.
  - Declassifying the yes and no outcomes when some discrimination criterion is varied.
  - In this case, the criterion is a relative misclassification cost.

Source: CognitiveClass

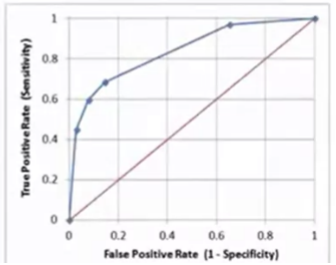
Case Study - 8. Evaluation

We can see that model 3, with a relative misclassification cost of 4-to-1, is the best of the 4 models

Case Study – Using the ROC curve

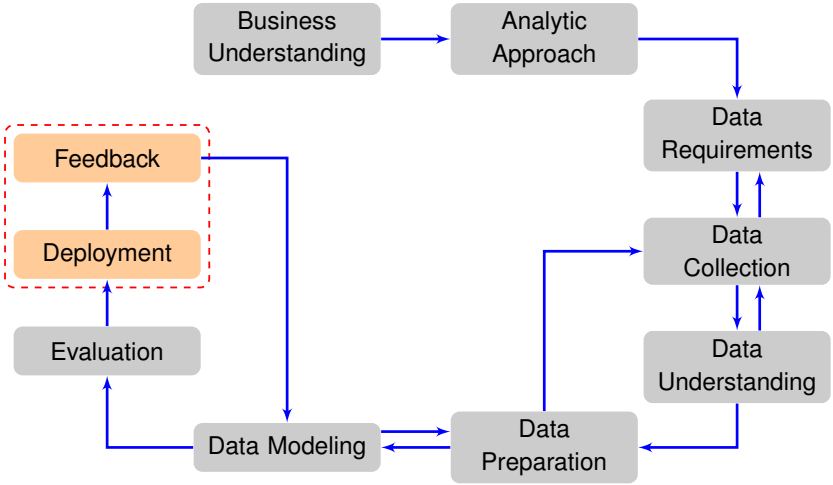
Diagnostic tool for classification model evaluation

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



Source: CognitiveClass

From Deployment to Feedback



From Deployment to Feedback

- The importance of stakeholder input.
- To consider the scale of deployment.
- The importance of incorporating feedback to refine the model.
- The refined model must be redeployed.
- This process should be repeated as often as necessary.

Source: CognitiveClass

9. Deployment (Concept)

- To make the model relevant and useful to address the initial question, involves getting the stakeholders familiar with the tool produced.
- Once the model is evaluated/approved by the stakeholders, it is deployed and put to the ultimate test.
- The model may be rolled out to a limited group of users or in a test environment, to build up confidence in applying the outcome for use across the board.

Source: CognitiveClass

Case Study - 9. Deployment

- Understanding the results
  - In preparation for model deployment, the next step was to assimilate the knowledge for the business group who would be designing and managing the intervention program to reduce readmission risk.
  - In this scenario, the business people translated the model results so that the clinical staff could understand how to identify high-risk patients and design suitable intervention actions.
  - The goal was to reduce the likelihood that these patients would be readmitted within 30 days after discharge.
  - During the business requirements stage, the Intervention Program Director and her team had wanted an application that would provide automated, near real-time risk assessments of congestive heart failure.

Case Study - 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

Source: CognitiveClass

Case Study - 9. Deployment

- Gathering application requirements
  - It also had to be easy for clinical staff to use, and preferably through browser-based application on a tablet, that each staff member could carry around.
  - This patient data was generated throughout the hospital stay.
  - It would be automatically prepared in a format needed by the model and each patient would be scored near the time of discharge.
  - Clinicians would then have the most up-to-date risk assessment for each patient, helping them to select which patients to target for intervention after discharge. As part of solution deployment, the Intervention team would develop and deliver training for the clinical staff.

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

Source: CognitiveClass

Case Study - 9. Deployment

- Additional Requirements
  - Processes for tracking and monitoring patients receiving the intervention would have to be developed in collaboration with IT developers and database administrators, so that the results could go through the feedback stage and the model could be refined over time.

Source: CognitiveClass

10. Feedback (Concept)

- Feedback from users to refine the model.
- Assess the model for performance and impact.
- The value of the model will be dependent on successfully incorporating feedback and making adjustments for as long as the solution is required.
- Throughout the Data Science Methodology, each step sets the stage for the next.
- This makes the methodology cyclical, ensures refinement at each stage in the game.
- Once the model has been evaluated and the data scientist trusts that it will work, it will be deployed and will undergo the final test:
  Its real use in real time in the field.

Source: CognitiveClass

Case Study - 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

Source: CognitiveClass

Case Study - 10. Feedback

Feedback stage included these steps:

1. The review process would be defined and put into place, with overall responsibility for measuring the results of the model applied to CHF risk population. Clinical management executives would have overall responsibility for the review process.
2. CHF patients receiving intervention would be tracked and their re-admission outcomes recorded.
3. The intervention would then be measured to determine how effective it was in reducing readmissions.

Source: CognitiveClass

Case Study - 10. Feedback

- For ethical reasons, CHF patients would not be split into controlled and treatment groups.
- Instead, readmission rates would be compared before and after the implementation of the model to measure its impact.
- After the deployment and feedback stages, the impact of the intervention program on re-admission rates would be reviewed after the first year of its implementation.
- Then the model would be refined, based on all of the data compiled after model implementation and the knowledge gained throughout these stages.

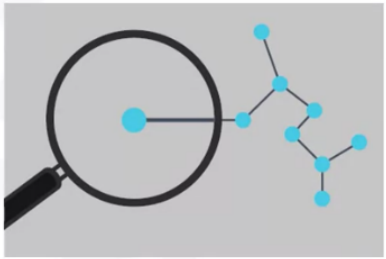
Source: CognitiveClass

Case Study - 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



Source: CognitiveClass

Case Study - 10. Feedback

Redeployment

- The intervention actions and processes would be reviewed and very likely refined as well, based on the experience and knowledge gained through initial deployment and feedback.
- Finally, the refined model and intervention actions would be redeployed, with the feedback process continued throughout the life of the Intervention program.

Source: CognitiveClass

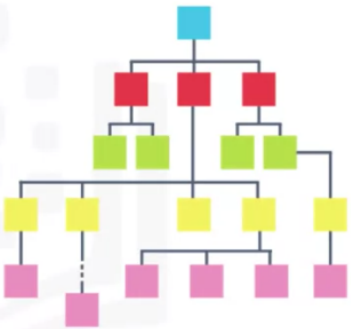
Case Study - 10. Feedback

Case Study – Redeployment

Review and refine intervention actions

Redeploy

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



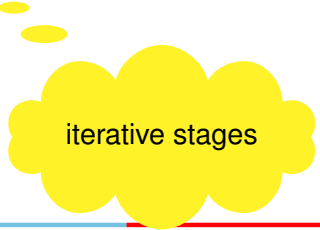
Source: CognitiveClass

Data Science Process - Summary

- 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

Data Science Process - Summary

- 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