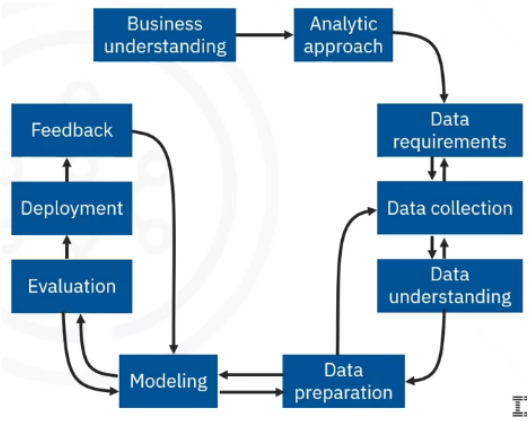
- Module 1: From Problem to Approach and From Requirements to Collection
- Module 2: From Understanding to Preparation and From Modeling to Evaluation
- Module 3: From Deployment to Feedback and Final Evaluation
- Module 4: Final Project and Assessment

❖ **MODULE 1: From Problem to Approach and From Requirements to Collection**

Data Science Methodology Overview

- Addressing data science challenges
    - Data Science combines statistics, technology, and domain expertise to extract insights from vast data.
    - Challenges:
        - Resolve the problems of misunderstanding of the business questions
        - Not knowing how to apply the data to resolve the business problem correctly
    - Adopting a methodology can help address these issues
- What is a methodology?
    - A system of methods
    - A guideline for decision-making during the scientific process
    - Data Science Methodology guides data scientists in solving complex problems with data
- Applying data science methodology
    - Perform data collection
    - Creation of measurement strategies
    - Comparisons of data analysis methods
- Addressing data science challenges
    - Apply practical guidance
    - Avoid the mistakes that can happen by jumping to solutions before the analysis
- Data methodology stages
    - 

    - Business understanding (define the issue)
        - What is the problem that you are trying to solve?
    - Analytic approach (determine your approach)
        - How can you use data to answer the business question?
    - Data requirements

        - What data do you need to answer the question?
    - Data collection
        - Where is the data sourced from, and how will you receive the data?
    - Data understanding
        - Does the data you collected represent the problem to be solved?
    - Data preparation
        - What additional work is required to manipulate and work with the data?
    - Modeling
        - When you apply data visualizations, do you see answers that address the business problem?
    - Evaluation
        - Does the data model answer the initial business question, or must you adjust the data?
    - Deployment
        - Can you put the model into practice?
    - Feedback
        - Can you get constructive feedback from the data and the stakeholder to answer the business question?

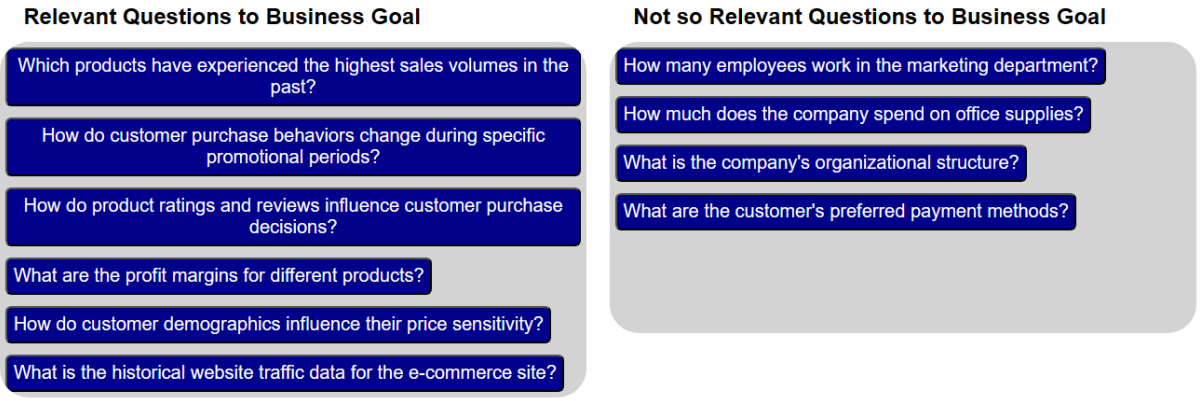
Business Understanding

- Case study
    - Goals and objectives
        - Define the GOALS
        - Define the OBJECTIVES
    - What's the sponsor's involvement?
        - Set overall direction
        - Remain engaged and provide guidance
        - Ensure necessary support, where needed

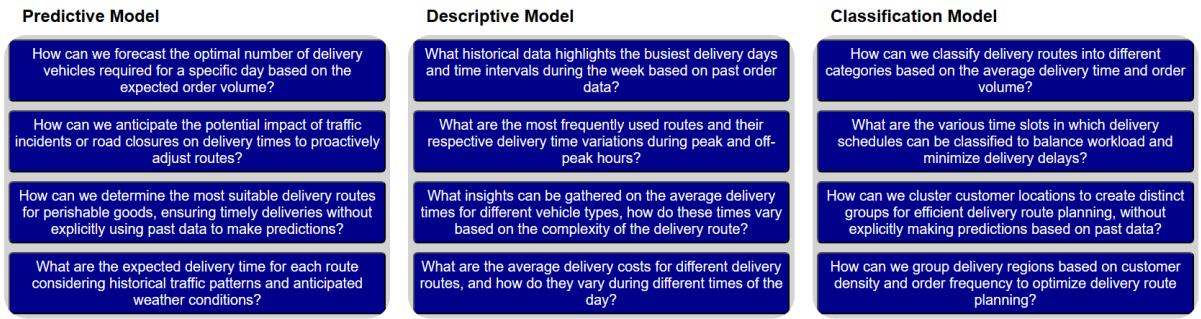
Analytic Approach

- Pick analytic approach based on the question type
    - 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?
- What are types of questions?
    - If the question is to determine the probabilities of an action
        - Use a predictive model
    - If the question is to show relationships
        - Use a descriptive model
    - If the question requires a yes/no answer
        - Use a classification model
- Which machine learning will be utilized
    - Learning without being explicitly programmed
    - Identifies relationships and trends in data that might otherwise not be accessible
    - Uses clustering association approaches
- Case study: Decision tree classification select
    - 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

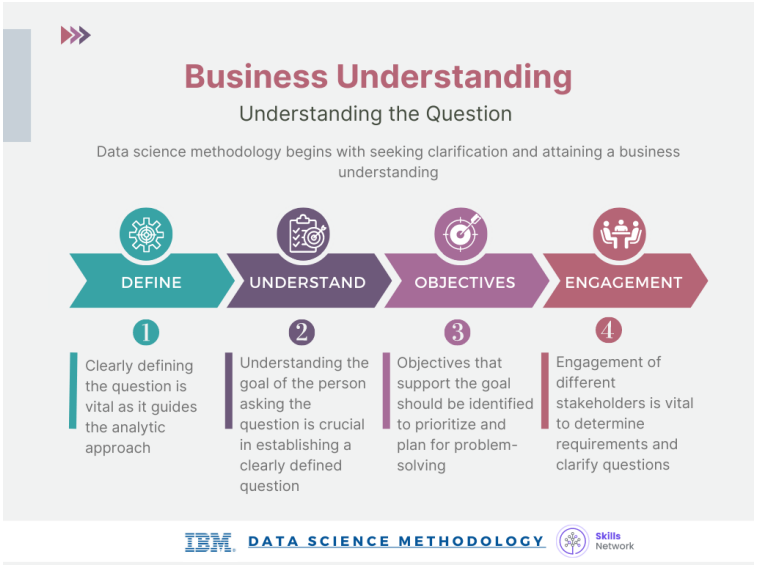


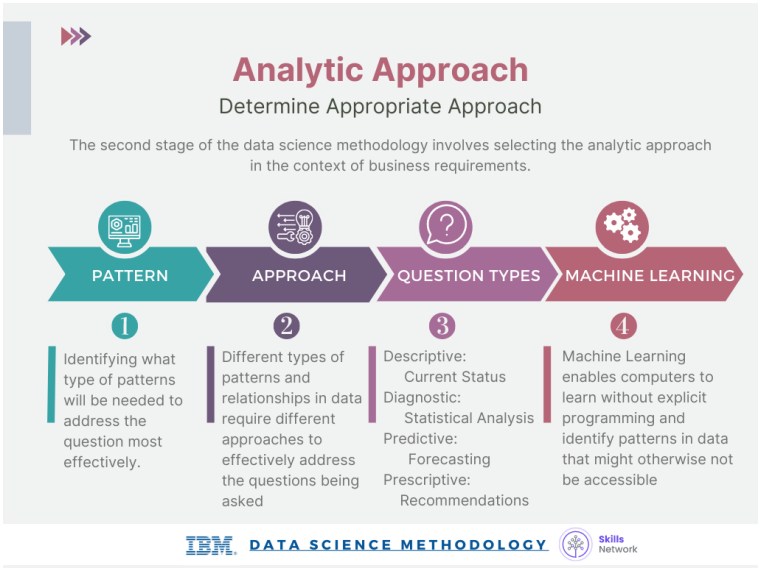
- Check your score Score 100 % Keep trying until you score 100 percent! You've got this! Start Over



- Check your score Score 100 % Keep trying until you score 100 percent! You've got this! Start Over

LESSON SUMMARY

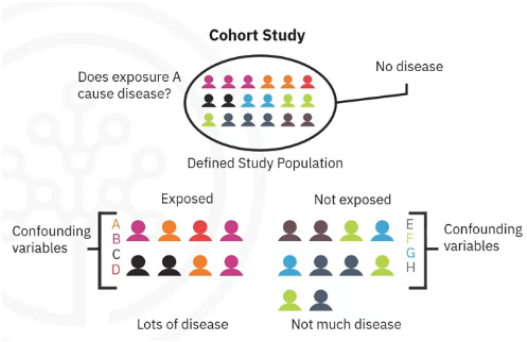


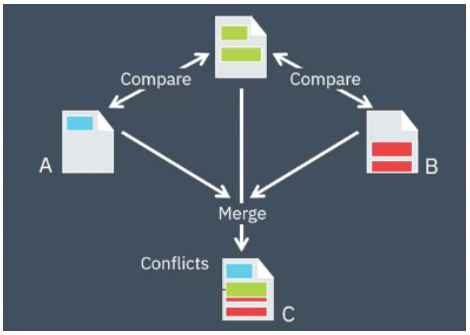


Glossary:

https://author-ide.skills.network/render?token=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJtZF9pbnN0cnVjdGlvbnNfdXJsIjoiaHR0cHM6Ly9jZi1jb3Vyc2VzLWRhdGEuczMudXMuY2xvdWQtb2JqZWN0LXN0b3JhZ2UuYXBwZG9tYWluLmNsb3VkL0lCTVNraWxsc05ldHdvcmstRFMwMTAzRU4tQ291cnNlcmEvbGFicy92NGdsb3NzYXJpZXMvTTFMMS5tZCIsInRvb2xfdHlwZSI6Imluc3RydWN0aW9uYWwtbGFiIiwiYWRtaW4iOmZhbHNlLCJpYXQiOjE2OTIyMDc1NTB9.tJnxmreG7T5mfOgkID0fw62LAkQpv23Zq4uoPHgUp8Q

Data Requirements

- Case Study: Selecting the cohort
  - In order to compile the complete clinical histories, three criteria were identified for inclusion in the cohort.
  - Inpatient 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 (excluded in the study in order to decrease the sample size of the patients included in the study.)
  - 

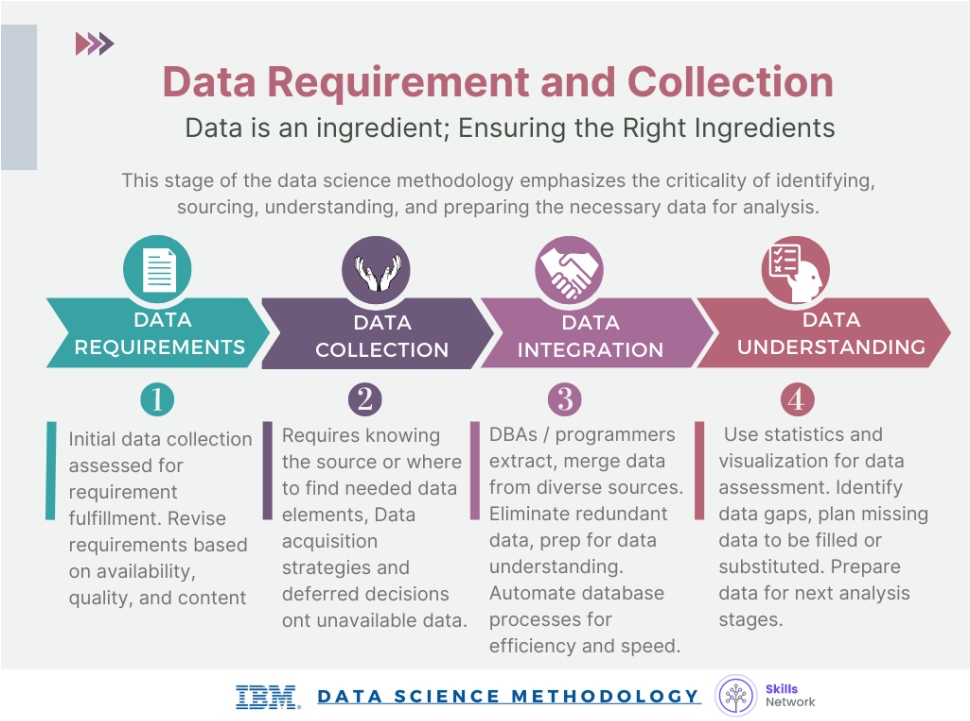
- Case Study: Defining the data
  - Content, formats, representations suitable for decision tree classifier
    - One record per patient with columns representing variables (dependent variable and predictors)
    - Content covering all aspects of each patient's clinical history
      - Transactional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

Data Collection

- Case Study: Gathering available data
  - Corporate data warehouse (single source of medical & claims, eligibility, provider, and member information)
  - Inpatient record system
  - Claim payment system
  - Disease management program information
- Case Study: Deferring inaccessible data

    - Data wanted but not available
        - Pharmaceutical records
        - Decided to defer
    - For this case study, certain drug information was also needed, but that data source was not yet integrated with the rest of the data sources. This leads to an important point: It is alright to defer decisions about unavailable data, and attempt to acquire it at a later stage.
- Case Study: Merging data
    - Eliminate redundant data
- Data scientists, essentially, explore the data to:
    - Understand its content
    - Assess its quality
    - Discover any interesting preliminary insights
    - Determine whether additional data is necessary to fill any gaps in the data

LESSON SUMMARY



Glossary:

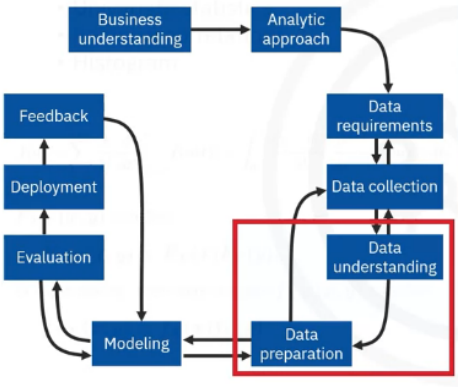
https://author-ide.skills.network/render?token=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJtZF9pbnN0cnVjdGlvbnNfdXJsIjoiaHR0cHM6Ly9jZi1jb3Vyc2VzLWRhdGEuczMudXMuY2xvdWQtb2JqZWN0LXN0b3JhZ2UuYXBwZG9tYWluLmNsb3VkL0lCTVNr

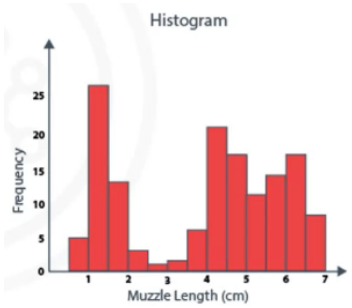
aWxsc05ldHdvcmstRFMwMTAzRU4tQ291cnNlcmEvbGFicy92NGdsb3NzYXJpZXMv
TTFMMi5tZCIsInRvb2xfdHlwZSI6Imluc3RydWN0aW9uYWwtbGFiIiwiYWRtaW4iOm
ZhbHNlLCJpYXQiOjE2OTIyMDUyMzV9.P_qUpQpI6Z2Pib9RYE8B1byi70jKE847ayk
qN1HgJjw

❖ **MODULE 2: From Understanding to Preparation and From Modeling to Evaluation**

Data Understanding

- Data Understanding: What does it mean to “prepare” or “clean” data?
- Data Preparation: What are ways in which data is prepared?

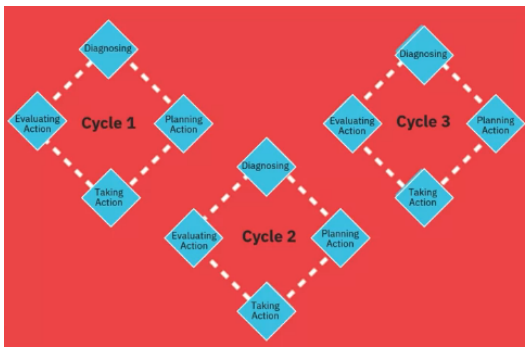


- Case Study (heart failure admissions): Understanding the Data
  - Descriptive statistics
    - Univariate statistics
    - Pairwise correlations
    - Histogram
  - Histograms are a good way to understand:
    - How values or variables are distributed
    - What data preparation might be needed to make the variable more useful in a model
  - 

  - First, these statistics included hurst, univariates, and statistics on each variable, such as mean, median, minimum, maximum, and standard deviation.
  - Second, pairwise correlations were used, to see how closely certain variables were related, and which ones, if any, were very highly correlated, meaning that they would be essentially redundant, thus making only one relevant for modeling.
  - Third, histograms of the variables were examined to understand their distributions. Histograms are a good way to understand how values or a variable are distributed, and which sorts of data preparation may be

needed to make the variable more useful in a model.

- Case Study: Looking at data quality
  - Missing values
  - Invalid or misleading values
  - 

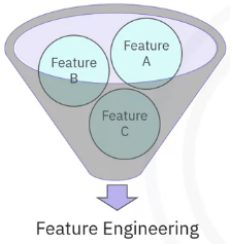
- Case Study: This is an iterative process
  - Iterative data collection and understanding
    - Refined definition of “CHF admission”
  - 

- The Data Understanding stage encompasses sorting the data.

Data Preparation - Concepts

- Cleansing data
  - In a sense, data preparation is similar to washing freshly picked vegetables insofar as unwanted elements, such as dirt or imperfections, are removed.
  - Together with data collection and data understanding, data preparation is the most time-consuming phase of a data science project, typically taking 70% and even up to even 90% of the overall project time.
  - Automating some of the data collection and preparation processes in the database, can reduce this time to as little as 50%. This time savings translates into increased time for data scientists to focus on creating models.
- Transforming data
  - Transforming data in the data preparation phase is the process of getting the data into a state where it may be easier to work with.
- Examples of data cleansing
  - To work effectively with the data, it must be prepared in a way that

addresses missing or invalid values and removes duplicates, toward ensuring that everything is properly formatted.

- Using domain knowledge
    - Feature engineering is the process of using domain knowledge of the data to create features of the data to create features that make the machine learning algorithms work
    - Feature engineering is critical when machine learning tools are being applied to analyze the data
    - 

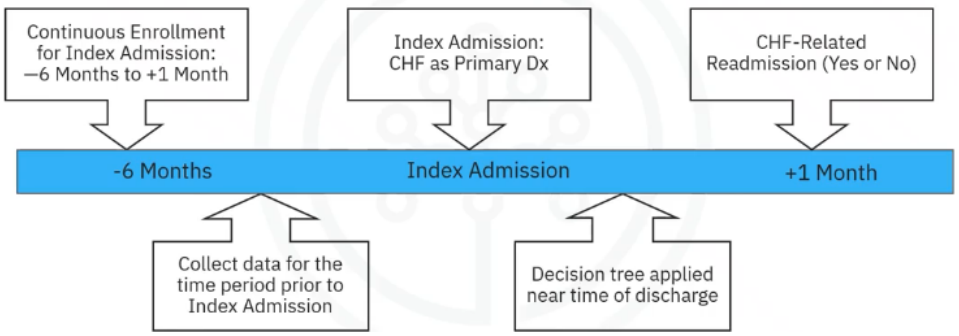
    - A feature is a characteristic that might help when solving a problem. Features within the data are important to predictive models and will influence the results you want to achieve. Feature engineering is critical when machine learning tools are being applied to analyze the data.
- Working with text analysis
    - When working with text, text analysis steps for coding the data are required to be able to manipulate the data. The data scientist needs to know what they're looking for within their dataset to address the question. The text analysis is critical to ensure that the proper groupings are set, and that the programming is not overlooking what is hidden within.
- The data preparation phase sets the stage for the next steps in addressing the question. While this phase may take a while to do, if done right the results will support the project. If this is skipped over, then the outcome will not be up to par and may have you back at the drawing board. It is vital to take your time in this area, and use the tools available to automate common steps to accelerate data preparation. Make sure to pay attention to the details in this area.

Data Preparation - Case Study

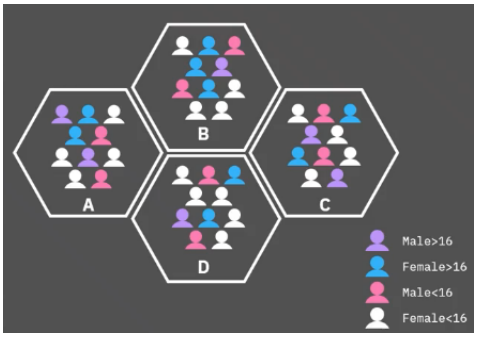
- In a sense, data preparation is similar to washing freshly picked vegetables insofar as unwanted elements, such as dirt or imperfections, are removed.
- Case Study: Data preparation
    - In the case study, an important first step in the data preparation stage was to actually define congestive heart failure. This sounded easy at first but defining it precisely was not straightforward. First, the set of diagnosis-related group codes needed to be identified, as congestive

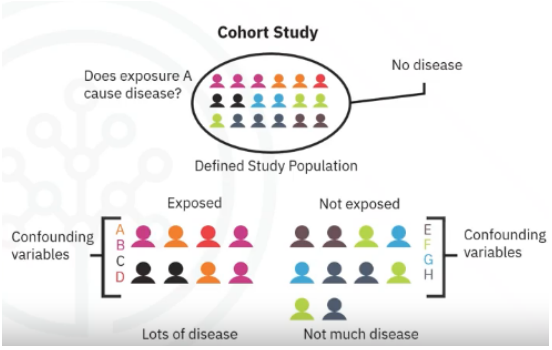
heart failure implies certain kinds of fluid buildup. We also needed to consider that congestive heart failure is only one type of heart failure. Clinical guidance was needed to get the right codes for congestive heart failure.

- The next step involved defining the re-admission criteria for the same condition. The timing of events needed to be evaluated in order to define whether a particular congestive heart failure admission was an initial event, which is called an index admission, or a congestive heart failure-related re-admission. Based on clinical expertise, a time period of 30 days was set as the window for readmission relevant for congestive heart failure patients, following the discharge from the initial admission.
- Next, the records that were in transactional format were aggregated, meaning that the data included multiple records for each patient.



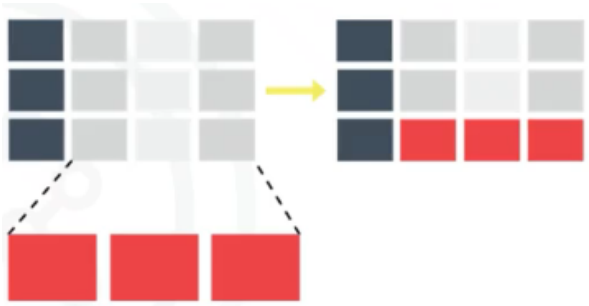
- Case study: Aggregating records (transactional records)
  - Claims: professional provider, facility, pharmaceutical
  - Inpatient & outpatient records: diagnoses, procedures, prescriptions and more
  - Possibly thousands per patient, depending on clinical history



- Case study: Aggregating to patient level
    - Then, all the transactional records were aggregated to the patient level, yielding a single record for each patient, as required for the decision-tree classification method that would be used for modeling. As part of the aggregation process, many new columns were created representing the information in the transactions.
    - 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
    - 

- Case study: More or less data needed?
    - Literature review of important factors of CHF readmission
    - Loop back to the data collection stage and add additional data, if needed
- Case study: Creating new variables
    - The result was the creation of one table containing a single record per patient, with many columns representing the attributes about the patient in his or her clinical history. These columns would be used as variables in the predictive modeling. Here is a list of the variables that were ultimately used in building the model. The dependent variable, or target, was congestive heart failure readmission within 30 days following discharge from a hospitalization for congestive heart failure, with an outcome of either yes or no.
    - The data preparation stage resulted in a cohort of 2,343 patients meeting all of the criteria for this case study. The cohort was then split into training and testing sets for building and validating the model, respectively.
    - 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, Age, Primary DRG, Length of stay, Prior admissions, Line of business, CHF Diagnosis importance (primary, secondary, tertiary)
- Diagnosis flags (Y/N)
  - CHF, Diabetes, Atrial fibrillation, Renal failure, Pneumonia, Hypertension

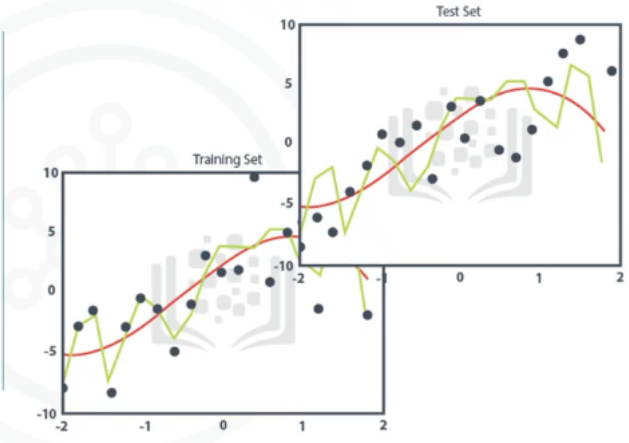


Cohort: 2,343 patients

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

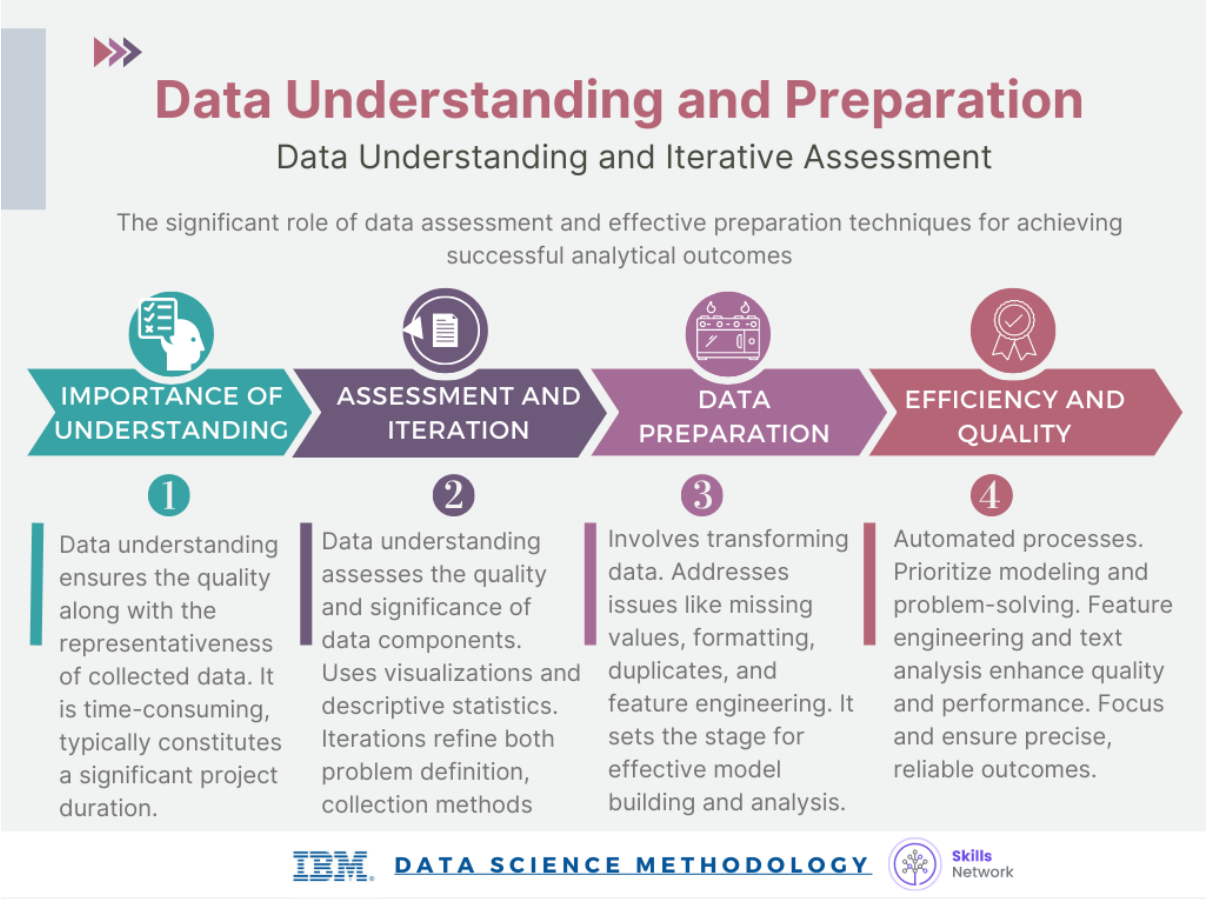
Training: 1,640 patients

Testing: 703 patients



LESSON SUMMARY

- The Data Understanding stage encompasses all activities related to constructing the data set and answers the question as to whether the data you collected represents the problem to be solved.
- During the Data Understanding stage, scientists might use descriptive statistics, predictive statistics, or both.
- Data scientists commonly apply Hurst, univariates, and other statistics on each variable, such as mean, median, minimum, maximum, standard deviation, pairwise correlation, and histograms.
- Data scientists also use univariates, statistics, and histograms to assess data quality.



- During the Data Preparation stage, data scientists must address missing or invalid values, remove duplicates, and validate that the data is properly formatted.
- Feature engineering, also part of the Data Preparation stage, uses domain knowledge of the data to create features that make the machine learning algorithms work.
- Text analysis during the Data Preparation stage is critical for validating that the proper groupings are set and that the programming is not overlooking hidden data.

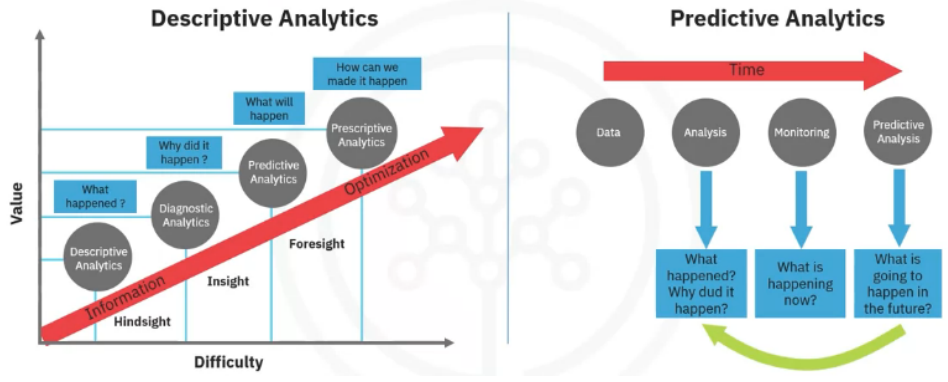
Glossary:

Automation	Using tools and techniques to streamline data collection and preparation processes.
Data Collection	The phase of gathering and assembling data from various sources.

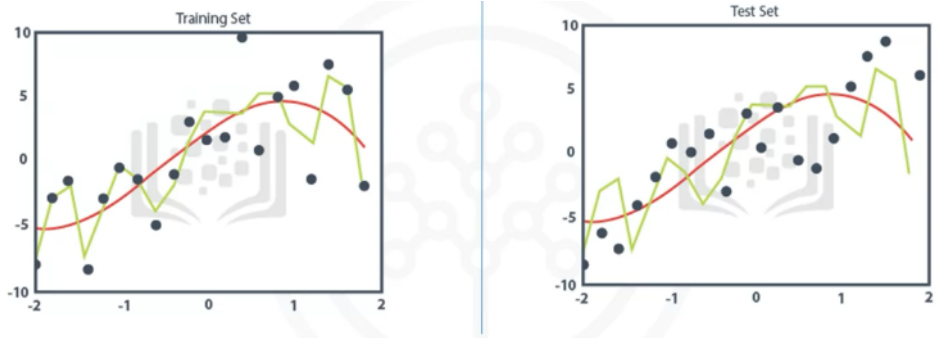
Data Compilation	The process of organizing and structuring data to create a comprehensive data set.
Data Formatting	The process of standardizing the data to ensure uniformity and ease of analysis.
Data Manipulation	The process of transforming data into a usable format.
Data Preparation	The phase where data is cleaned, transformed, and formatted for further analysis, including feature engineering and text analysis. The stage where data is transformed and organized to facilitate effective analysis and modeling.
Data Quality	Assessment of data integrity and completeness, addressing missing, invalid, or misleading values.
Data Quality Assessment	The evaluation of data integrity, accuracy, and completeness.
Data Set	A collection of data used for analysis and modeling.
Data Understanding	The stage in the data science methodology focused on exploring and analyzing the collected data to ensure that the data is representative of the problem to be solved.
Descriptive Statistics	Summary statistics that data scientists use to describe and understand the distribution of variables, such as mean, median, minimum, maximum, and standard deviation.
Feature	A characteristic or attribute within the data that helps in solving the problem.
Feature Engineering	The process of creating new features or variables based on domain knowledge to improve machine learning algorithms' performance.
Feature Extraction	Identifying and selecting relevant features or attributes from the data set.
Interactive Processes	Iterative and continuous refinement of the methodology based on insights and feedback from data analysis.
Missing Values	Values that are absent or unknown in the dataset,

	requiring careful handling during data preparation.
Model Calibration	Adjusting model parameters to improve accuracy and alignment with the initial design.
Pairwise Correlations	An analysis to determine the relationships and correlations between different variables.
Text Analysis	Steps to analyze and manipulate textual data, extracting meaningful information and patterns.
Text Analysis Groupings	Creating meaningful groupings and categories from textual data for analysis.
Visualization techniques	Methods and tools that data scientists use to create visual representations or graphics that enhance the accessibility and understanding of data patterns, relationships, and insights

Modeling - Concepts

- Modeling is the stage in the data science methodology, where the data scientist has the chance to sample the sauce and determine if it's bang on or in need of more seasoning!
- Data modeling: Using predictive or prescriptive?
    - An example of a descriptive model might examine things like: if a person did this, then they're likely to prefer that. A predictive model tries to yield yes/no, or stop/go type outcomes. These models are based on the analytic approach that was taken, either statistically driven or machine learning driven.
    - 

- Data modeling: Using training/test sets
    - A training dataset is an initial dataset that teaches the ML models to identify desired patterns or perform a particular task. A testing dataset is used to evaluate how effective the training was or how accurate the model is.

    - The data scientist will use a training set for predictive modeling. A training set is a set of historical data in which the outcomes are already known. The training set acts like a gauge to determine if the model needs to be calibrated. In this stage, the data scientist will play around with different algorithms to ensure that the variables in play are actually required.
    - 

Does the model need to be calibrated?
- Understand the question
    - The success of data compilation, preparation and modeling, depends on the understanding of the problem at hand, and the appropriate analytical approach being taken. The data supports the answering of the question, and like the quality of the ingredients in cooking, sets the stage for the outcome. Constant refinement, adjustments and tweaking are necessary within each step to ensure the outcome is one that is solid.
        - Understand the question at hand
        - Select an analytic approach or method to solve the problem
        - Obtain, understand, prepare, and model the data
    - The end goal is to move the data scientist to a point where a data model can be built to answer the question.
- Was the question answered?
    - In this stage of the methodology, model evaluation, deployment, and feedback loops ensure that the answer is near and relevant. This relevance is critical to the data science field overall, as it is a fairly new field of study, and we are interested in the possibilities it has to offer. The more people that benefit from the outcomes of this practice, the further the field will develop.

Modeling - Case Study

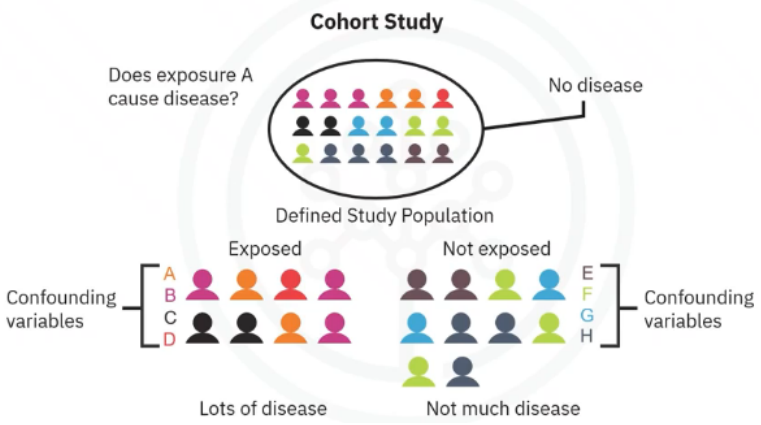
- Case Study: Analyzing the first model
    - With a prepared training set, the first decision tree classification model for congestive heart failure readmission can be built. We are looking for patients with high-risk readmission, so the outcome of interest will be congestive heart failure readmission equals "yes". In this first model,

overall accuracy in classifying the yes and no outcomes was 85%. This sounds good, but it represents only 45% of the "yes". The actual readmissions are correctly classified, meaning that the model is not very accurate. The question then becomes: How could the accuracy of the model be improved in predicting the yes outcome? For decision tree classification, the best parameter to adjust is the relative cost of misclassified yes and no outcomes.

- 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%

- Case study: How to improve the model?
  - Think of it like this: 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. A statistician calls this a type I error, or a false-positive. But when a true readmission is misclassified, and no action is taken to reduce that risk, then the cost of that error is the readmission and all its attended costs, plus the trauma to the patient.
  - This is a type II error, or a false-negative. So we can see that the costs of the two different kinds of misclassification errors can be quite different. For this reason, it's reasonable to adjust the relative weights of misclassifying the yes and no outcomes. The default is 1-to-1, but the decision tree algorithm allows the setting of a higher value for yes.

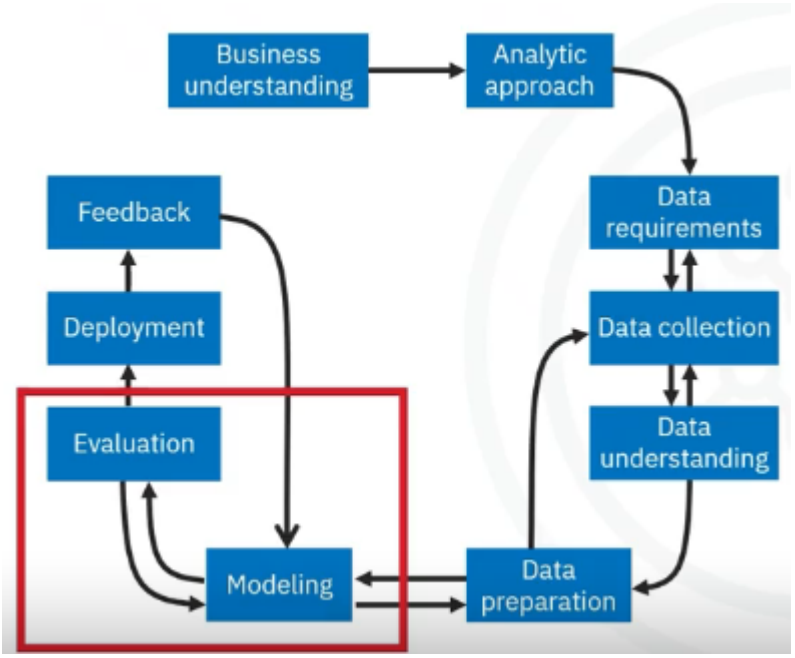


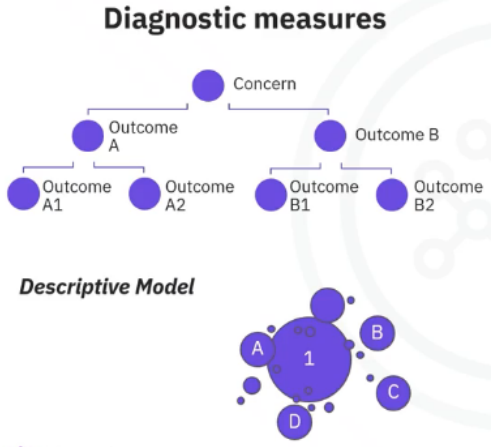
- Case study: Analyzing the second model
  - For the second model, the relative cost was set at 9-to-1. This is a very high ratio, but gives more insight to the model's behavior. This time the model correctly classified 97% of the yes, but at the expense of a very low accuracy on the no, with an overall accuracy of only 49%.

    - This was clearly not a good model. The problem with this outcome is the large number of false-positives, which would recommend unnecessary and costly intervention for patients, who would not have been re-admitted anyway. Therefore, the data scientist needs to try again to find a better balance between the yes and no accuracies.
- Case study: Analyzing the third model
    - For the third model, the relative cost was set at a more reasonable 4-to-1. This time 68% accuracy was obtained on only yes, called sensitivity by statisticians, and 85% accuracy on the no, called specificity, with an overall accuracy of 81%. This is the best balance that can be obtained with a rather small training set through adjusting the relative cost of misclassified yes and no outcomes parameters.
    - A lot more work goes into the modeling, of course, including iterating back to the data preparation stage to redefine some of the other variables, so as to better represent the underlying information, and thereby improve the model.

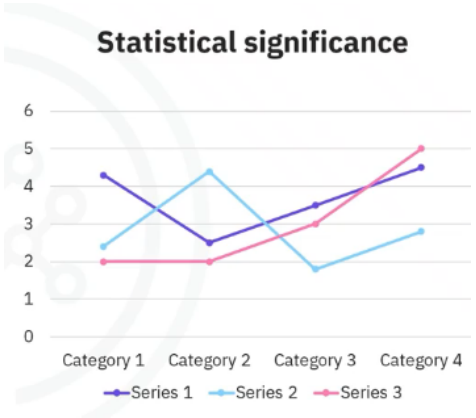
Evaluation

- A model evaluation goes hand-in-hand with model building as such, the modeling and evaluation stages are done iteratively.
- From modeling to evaluation
    - Model evaluation is performed during model development and before the model is deployed. Evaluation allows the quality of the model to be assessed but it's also an opportunity to see if it meets the initial request. Evaluation answers the question: Does the model used really answer the initial question or does it need to be adjusted?
    - Modeling
        - In what way can the data be visualized to get to the answer that is required
    - Evaluation
        - Initial question or does it need to be adjusted?

- 

- When and not to adjust the model?
  - Model evaluation can have two main phases. The first is the diagnostic measures phase, which is used to ensure the model is working as intended. If the model is a predictive model, a decision tree can be used to evaluate if the answer the model can output is aligned to the initial design. It can be used to see where there are areas that require adjustments. If the model is a descriptive model, one in which relationships are being assessed, then a testing set with known outcomes can be applied, and the model can be refined as needed.
  - The second phase of evaluation that may be used is statistical significance testing. This type of evaluation can be applied to the model to ensure that the data is being properly handled and interpreted within the model. This is designed to avoid unnecessary second guessing when the answer is revealed.
  - Diagnostic measures
  - 

  - Statistical significance

- 

- Case Study: Applying the concepts
  - Let's look at one way 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. As shown in this table, four models were built with four different relative misclassification costs
Model	Relative Cost Y:N	True Positive Rate (Sensitivity)	Specificity (accuracy in 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

  - Misclassification cost tuning
    - Tune the relative misclassification costs
    - Balance true-positive rate and false-positive rate for best model
- Case Study: Relative costs
  - As we see, each value of this model-building parameter increases the true-positive rate, or sensitivity, of the accuracy in predicting yes, at the expense of lower accuracy in predicting no, that is, an increasing false-positive rate. The question then becomes, which model is best based on tuning this parameter? For budgetary reasons, the risk-reducing intervention could not be applied to most or all congestive heart failure patients, many of whom would not have been readmitted anyway.
Model	Relative Cost Y:N
1	1:1
2	1.5:1
3	4:1
4	9:1

- Case Study: True-positives vs false-positives

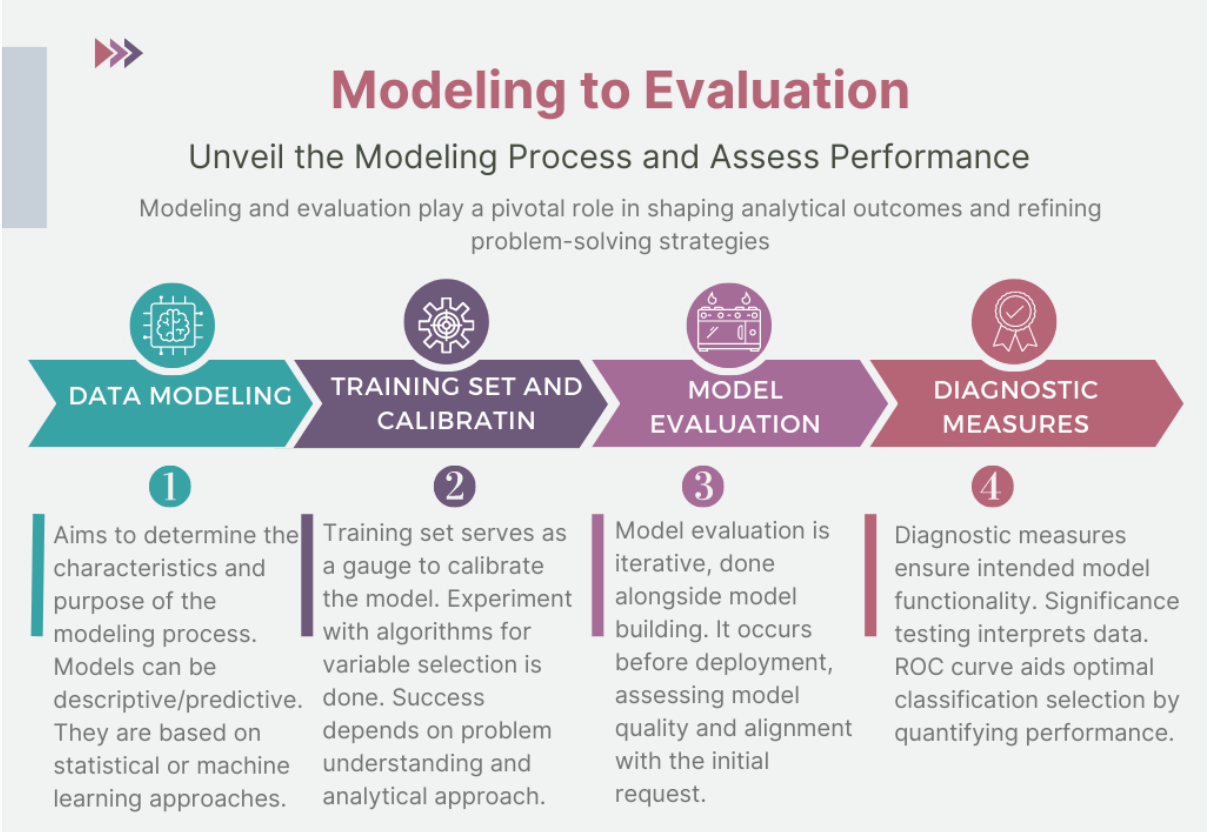
- On the other hand, the intervention would not be as effective in improving patient care as it should be, with not enough high-risk congestive heart failure patients targeted. So, how do we determine which model was optimal?
-

Relative Cost Y:N	True Positive Rate (Sensitivity)
1:1	0.45
1.5:1	0.60
4:1	0.68
9:1	0.97

- Case Study: Using the ROC curve
  - As you can see on this slide, the optimal model is the one giving the maximum separation between the blue ROC curve relative to the red baseline. We can see that model 3, with a relative misclassification cost of 4-to-1, is the best of the 4 models.
  - And just in case you were wondering, ROC stands for receiver operating characteristic curve, which was first developed during World War II to detect enemy aircraft on radar. It has since been used in many other fields as well. Today it is commonly used in machine learning and data mining.
  - The ROC curve is a useful diagnostic tool in determining the optimal classification model. 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. By plotting the true-positive rate against the false-positive rate for different values of the relative misclassification cost, the ROC curve helped in selecting the optimal model.
  - Diagnostic tool for classification model evaluation
    - Classification model performance
    - True-Positive Rate vs False-Positive Rate
    - Optimal model at maximum separation
  - 

LESSON SUMMARY

- The end goal of the Modeling stage is that the data model answers the business question.
- The data modeling process uses a training data set. Data scientists test multiple algorithms on the training set data to determine whether the variables are required and whether the data supports answering the business question. The outcome of those models are either descriptive or predictive.



- The Evaluation phase consists of two stages, the diagnostic measures phase, and the statistical significance phase.
- During the Evaluation stage, data scientists and others assess the quality of the model and determine if the modelanswers the initial Business Understanding question or if the data model needs adjustment.
- The ROC curve, known as the receiver operating characteristic curve, is a useful diagnostic tool for determining the optimal classification model. This curve quantifies how well a binary classification model performs, declassifying the yes and no outcomes when some discrimination criterion is varied.

Glossary:

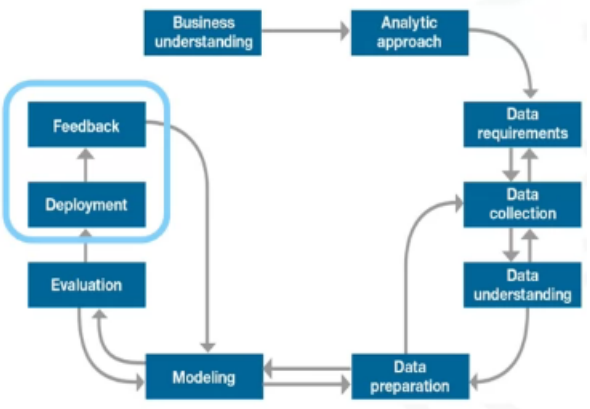
Binary classification	A model that classifies data into two categories, such

model	as yes/no or stop/go outcomes.
Data compilation	The process of gathering and organizing data required for modeling.
Data modeling	The stage in the data science methodology where data scientists develop models, either descriptive or predictive, to answer specific questions.
Descriptive model	A type of model that examines relationships between variables and makes inferences based on observed patterns.
Diagnostic measure based tuning	The process of fine-tuning the model by adjusting parameters based on diagnostic measures and performance indicators.
Diagnostic measures	The evaluation of a model's performance to ensure that the model functions as intended.
Discrimination criterion	A measure used to evaluate the performance of the model in classifying different outcomes.
False-positive rate	The rate at which the model incorrectly identifies negative outcomes as positive.
Histogram	A graphical representation of the distribution of a dataset, where the data is divided into intervals or bins, and the height of each bar represents the frequency or count of data points falling within that interval.
Maximum separation	The point where the ROC curve provides the best discrimination between true-positive and false-positive rates, indicating the most effective model.
Model evaluation	The process of assessing the quality and relevance of the model before deployment.
Optimal model	The model that provides the maximum separation between the ROC curve and the baseline, indicating higher accuracy and effectiveness.
Receiver Operating Characteristic (ROC)	Originally developed for military radar, this statistical curve is used to assess the performance of binary classification models.
Relative misclassification	This measurement is a parameter in model building

cost	used to tune the trade-off between true-positive and false-positive rates.
ROC curve (Receiver Operating Characteristic curve)	A diagnostic tool used to determine the optimal classification model's performance.
Separation	Separation is the degree of discrimination achieved by the model in correctly classifying outcomes.
Statistical significance testing	Evaluation technique to verify that data is appropriately handled and interpreted within the model.
True-positive rate	The rate at which the model correctly identifies positive outcomes.

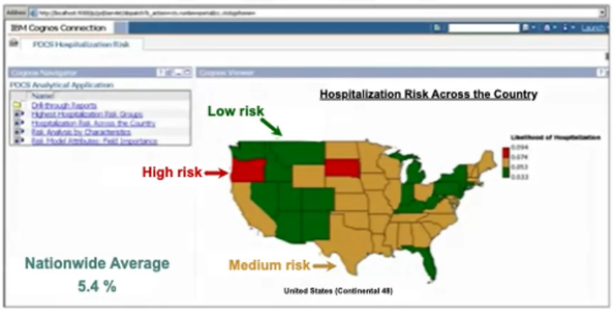
❖ MODULE 3: From Deployment to Feedback and Final Evaluation

Deployment

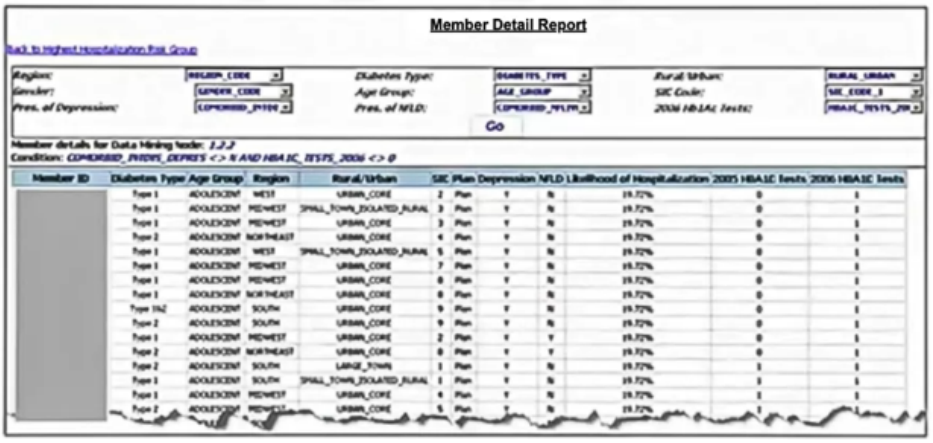
- Are the stakeholders familiar with the new tool?
    - While a data science model will provide an answer, the key to making the answer relevant and useful to address the initial question, involves getting the stakeholders familiar with the tool produced. In a business scenario, stakeholders have different specialties that will help make this happen, such as the solution owner, marketing, application developers, and IT administration.
- From deployment to feedback
    - Once the model is evaluated and the data scientist is confident it will work, it is deployed and put to the ultimate test. Depending on the purpose of the model, it 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
    - 

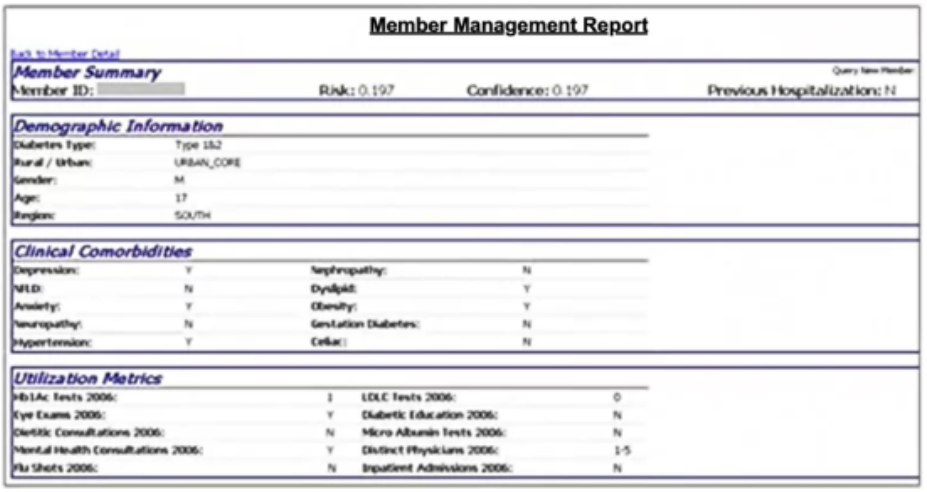
- Case study: Understand the results
    - In preparation for solution deployment, the next step was to assimilate the knowledge for the business group who would be designing and managing the intervention program to reduce readmission risk. In this scenario, the business people translated the model results so that the clinical staff could understand how to identify high-risk patients and design suitable intervention actions. The goal, of course, was to reduce the likelihood that these patients would be readmitted within 30 days after discharge.
    - Assimilate knowledge for business:
        - Practical understanding of the meaning of model results
        - Implications of model results for designing intervention actions
- Case study: Gathering application requirements
    - 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.

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.

    - 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
- Case study: Additional requirements?
    - As part of solution deployment, the Intervention team would develop and deliver training for the clinical staff. Also, 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.
    - Additional requirements:
        - Training for clinical staff
        - Tracking / monitoring processes
- Example 1: Solution deployment
    - This map is an example of a solution deployed through a Cognos application. In this case, the case study was hospitalization risk for patients with juvenile diabetes. Like the congestive heart failure use case, this one used decision tree classification to create a risk model that would serve as the foundation for this application. The map gives an overview of hospitalization risk nationwide, with an interactive analysis of predicted risk by a variety of patient conditions and other characteristics.
    - Hospitalization risk for juvenile diabetes patients
    - 

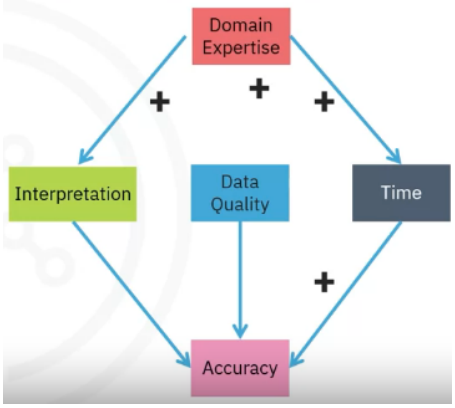
- Example 2: Solution deployment

- This slide shows an interactive summary report of risk by patient population within a given node of the model, so that clinicians could understand the combination of conditions for this subgroup of patients.
- Risk summary report by decision tree model note
- 

- Example 3: Solution deployment
  - And this report gives a detailed summary on an individual patient, including the patient's predicted risk and details about the clinical history, giving a concise summary for the doctor.
  - Individual patient risk report
  - 

Feedback

- Feedback: Problem solved? Question answered?
  - Once in play, feedback from the users will help to refine the model and assess it 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. Making the methodology cyclical, ensures refinement at each stage in the game. The feedback process is rooted in the notion that, the more you know, the more that you'll want to know.

- From deployment to feedback
    - Once the model is evaluated and the data scientist is confident it will work, it is deployed and put to the ultimate test:
        - Actual real-time use in the field
- Case study: Assessing model performance
    - First, the review process would be defined and put into place, with overall responsibility for measuring the results of a "flying to risk" model of the congestive heart failure risk population. Clinical management executives would have overall responsibility for the review process.
    - Second, congestive heart failure patients receiving intervention would be tracked and their re-admission outcomes recorded.
    - Third, the intervention would then be measured to determine how effective it was in reducing readmissions. For ethical reasons, congestive heart failure 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.
    - 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
    - 

- Case study: Refinement
    - After the deployment and feedback stages, the impact of the intervention program on readmission 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. Other refinements included: Incorporating information about participation in the intervention program, and possibly refining the model to incorporate detailed pharmaceutical data.

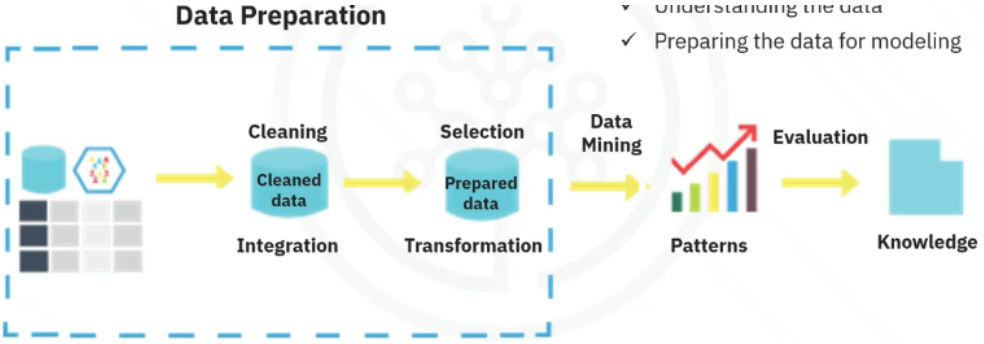
        - If you recall, data collection was initially deferred because the pharmaceutical data was not readily available at the time. But after feedback and practical experience with the model, it might be determined that adding that data could be worth the investment of effort and time. We also have to allow for the possibility that other refinements might present themselves during the feedback stage.
        - 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
- Case study: Redeployment
    - Also, 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.
    - Review and refine intervention actions
    - Redeploy
        - Continue modeling, deployment, feedback, and refinement throughout, and refinement throughout the life of the intervention program

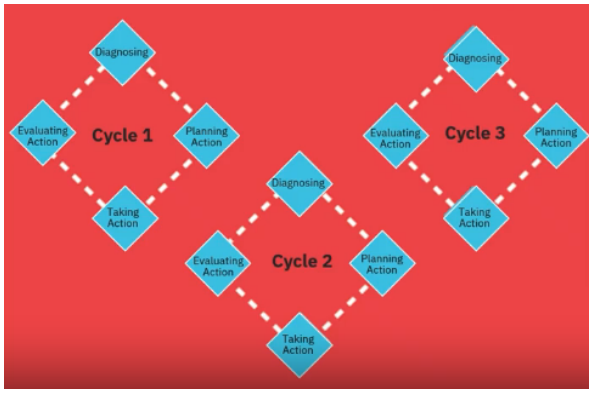
Storytelling

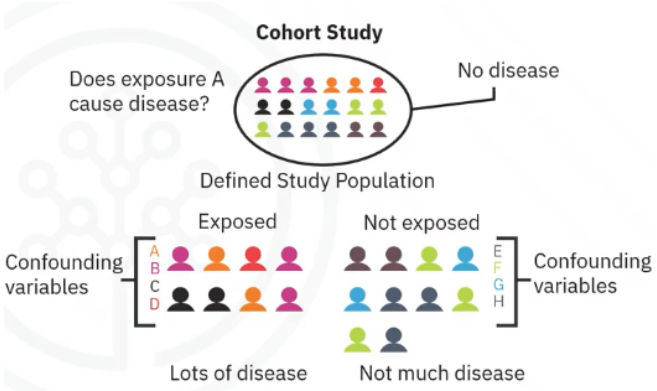
- What role does storytelling play in Data Analysis?
    - Storytelling with data is a critical skill for Data Analysts
    - It's important to tell a clear, concise, and compelling story to convince people to take action
    - Develop a story for your data set to understand your data better
    - Find the balance between telling a simple story and conveying the complexities of the data
    - It doesn't matter what information you have if you can't communicate it effectively to your audience
    - The best way to communicate your information is through visuals and telling a story
    - Storytelling is an essential skill set, the last mile in delivery
    - The ability to extract value from data and to tell a compelling story with data is critical
    - Storytelling is crucial to data analytics
    - Stories is how you convey your message
    - A compelling story helps your audience resonate with your findings

    - People remember stories
    - Stories help build an emotional connect and drive people to action

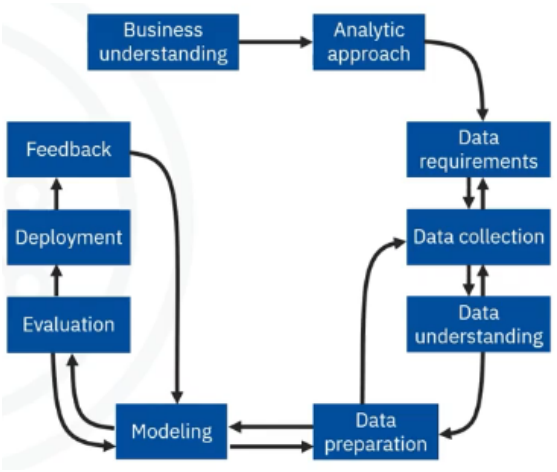
Course Summary

- From problem to approach
    - Thinking like a data scientist!
        - Forming a concrete business or research problem
        - Collecting and analyzing data
        - Building a model
        - Understanding the feedback after deployment
    - Learned the importance of
        - Understanding the question
        - Picking the most effective analytic approach
- To working with the data
    - Learned to work with data!
        - Determining the requirements
        - Collecting the appropriate data
        - Understanding the data
        - Preparing the data for modeling
    - 

- To deriving the answer
    - Once the analytic approach is selected, learn how to derive the answer
        - Evaluating and deploying the model
        - Getting feedback on it
        - Using that feedback constructively so as to improve the model
    - Remember that the stages of this methodology are iterative!
    - 

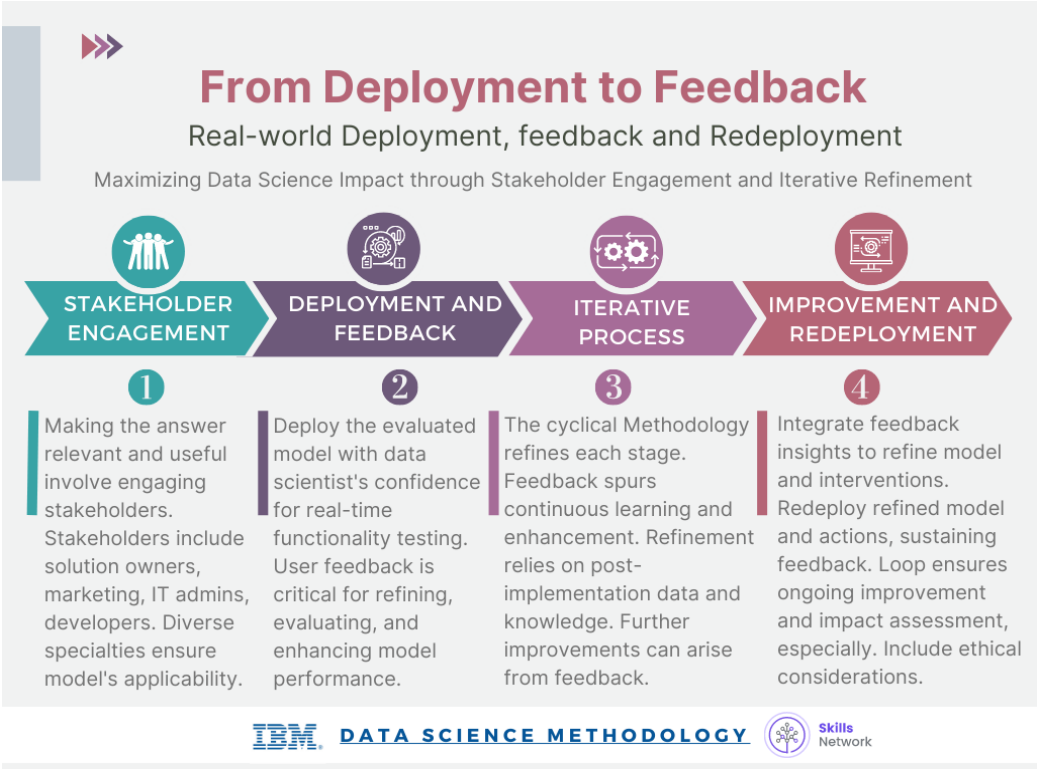
- Applied the concepts using a case study
    - Using a real case study, you learned how data science methodology can be applied in context, toward successfully achieving the goals that were set out in the business requirements stage. You also saw how the methodology contributed additional value to business units by incorporating data science practices into their daily analysis and reporting functions. The success of this new pilot program that was reviewed in the case study was evident by the fact that physicians were able to deliver better patient care by using new tools to incorporate timely data-driven information into patient care decisions.
    - 

- The methodology in a nutshell
    - The Data Science Methodology aims to answer the following 10 questions in this prescribed sequence:
    - From problem to approach:
        - 1. What is the problem that you are trying to solve?
        - 2. How can you use the data to answer the question?
    - Working with the data:
        - 3. What data do you need to answer the question?
        - 4. Where is the data coming from (identify all sources), and how will you get it?
        - 5. Is the data that you collected representative of the problem to be solved?
        - 6. What additional work is required to manipulate and work with the data
    - Deriving 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?



MODULE 3 SUMMARY

- Stakeholders, including the solution owner, marketing staff, application developers, and IT administration evaluate the model and contribute feedback.
- During the Deployment stage, data scientists release the data model to a targeted group of stakeholders.
- Stakeholder and user feedback help assess the model's performance and impact during the Feedback stage.
- The model's value depends on iteration; that is, how successfully the data model incorporates user feedback.



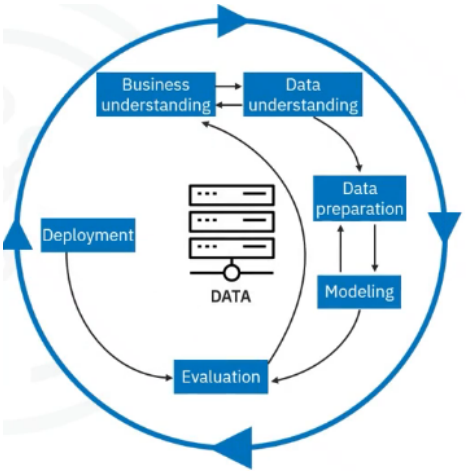
Glossary:

Browser-based application	An application that users access through a web browser, typically on a tablet or other mobile device, to provide easy access to the model's insights.
Cyclical methodology	An iterative approach to the data science process, where each stage informs and refines the subsequent stages.
Data collection refinement	The process of obtaining additional data elements or information to improve the model's performance.
Data science model	The result of data analysis and modeling that provides answers to specific questions or problems.
Feedback	The process of obtaining input and comments from users and stakeholders to refine and improve the data science model.
Model refinement	The process of adjusting and improving the data science model based on user feedback and real-world performance.
Redeployment	The process of implementing a refined model and intervention actions after incorporating feedback and improvements.
Review process	The systematic assessment and evaluation of the data science model's performance and impact.
Solution deployment	The process of implementing and integrating the data science model into the business or organizational workflow.
Solution owner	The individual or team responsible for overseeing the deployment and management of the data science solution.
Stakeholders	Individuals or groups with a vested interest in the data science model's outcome and its practical application, such as solution owners, marketing,

	application developers, and IT administration.
Storytelling	Storytelling is the art of conveying your message, or ideas through a narrative structure that engages, entertains, and resonates with the audience.
Test environment	A controlled setting where the data science model is evaluated and refined before full-scale implementation.

❖ **MODULE 4: Final Project and Assessment**

Introduction to CRISP-DM

- An acronym for Cross-Industry Standard Process for Data Mining
- A structured approach to guide data-driven decision-making
- The CRISP-DM model includes:
    - Data mining stages
    - Data mining stage descriptions
    - Explanations of the relationships between tasks and stages
    - 

- CRISP-DM: A high level process model
    - Provides high-level insights into the data mining life cycle
- Flexibility and communication using CRISP-DM
    - Data scientists might need to
        - Communicate with peers, management, and stakeholders to keep the project on track
        - Revisit earlier stages
- The Business Understanding stage:
    - Sets and outlines the project's data analysis intentions and goals
    - Requires communication and clarity to overcome stakeholders' differing objectives, biases, and information modalities
    - Is necessary to avoid wasted time and resources
- The Data Understanding stage:
    - CRISP-DM combines the stages of Data Requirements, Data Collection, and Data Understanding
    - Data scientists decide on data sources and acquire data
- The Data Preparation stage
    - Data scientists perform the following tasks:
        - Transform data
        - Determine if more data is needed
        - Address questionable missing and ambiguous data values
- The Modeling stage

    - Data mining:
        - Reveals patterns and structure within the data
        - Provides knowledge and insights that address the stated business problem and goals
    - Data scientists perform the following tasks:
        - Select data models
        - Adjust the models
- The Evaluation stage
    - Data scientists perform the following tasks:
        - Test the selected module
        - Assess the model's effectiveness
        - Results determine the model's efficacy
- The Deployment stage
    - Data scientists and stakeholders perform the following tasks:
        - Use the data model on new data outside of the data set
        - Analyze the results to determine the need for new variables, a new data set, or a new model
- CRISP-DM: Iterative and cyclical
    - Deployment results might initiate revisions to the following data analysis items
        - The business needs (question)
        - The necessary business actions
        - The data model
        - The data
- Discussing the results
    - After completing all six stages
        - Meet with the stakeholders to discuss the results
        - This stage is unnamed in CRISP-DM
        - This stage is the Feedback stage in John Rollins Foundational Data Science Model

REVIEW WHAT YOU LEARNED

- Foundational methodology, a cyclical, iterative data science methodology developed by John Rollins, consists of 10 stages, starting with Business Understanding and ending with Feedback.
- CRISP-DM, an open source data methodology, combines several data-related methodology stages into one stage and omits the Feedback stage resulting in a six-stage data methodology.
- The primary goal of the Business Understanding stage is to understand the business problem and determine the data needed to answer the core business question.

- During the Analytic Approach stage, you can choose from descriptive diagnostic, predictive, and prescriptive analytic approaches and whether to use machine learning techniques.
- During the Data Requirements stage, scientists identify the correct and necessary data content, formats, and sources needed for the specific analytical approach.
- During the Data Collection stage, expert data scientists revise data requirements and make critical decisions regarding the quantity and quality of data. Data scientists apply descriptive statistics and visualization techniques to thoroughly assess the content, quality, and initial insights gained from the collected data, identify gaps, and determine if new data is needed, or if they should substitute existing data.
- The Data Understanding stage encompasses all activities related to constructing the data set. This stage answers the question of whether the collected data represents the data needed to solve the business problem. Data scientists might use descriptive statistics, predictive statistics, or both.
- Data scientists commonly apply Hurst, univariates, and statistics such as mean, median, minimum, maximum, standard deviation, pairwise correlation, and histograms.
- During the Data Preparation stage, data scientists must address missing or invalid values, remove duplicates, and validate that the data is properly formatted. Feature engineering and text analysis are key techniques data scientists apply to validate and analyze data during the Data Preparation stage.
- The end goal of the Modeling stage is that the data model answers the business question. During the Modeling stage, data scientists use a training data set. Data scientists test multiple algorithms on the training set data to determine whether the variables are required and whether the data supports answering the business question. The outcome of those models is either descriptive or predictive.
- The Evaluation stage consists of two phases, the diagnostic measures phase, and the statistical significance phase. Data scientists and others assess the quality of the model and determine if the model answers the initial Business Understanding question or if the data model needs adjustment.
- During the Deployment stage, data scientists release the data model to a targeted group of stakeholders, including solution owners, marketing staff, application developers, and IT administration.,
- During the Feedback stage, stakeholders and users evaluate the model and contribute feedback to assess the model's performance.
- The data model's value depends on its ability to iterate; that is, how successfully the data model incorporates user feedback.