Module 1: From Problem to Approach and from Requirements to Collection

- Business Understanding
- Analytic Approach
- Data Requirements
- Data Collection
- Lab: From Problem to Approach
- Lab: From Requirement to Collection
- Quiz: From Problem to Approach
- Quiz: From Requirement to Collection

Module 2: From Understanding to Preparation and from Modeling to Evaluation

- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Lab: From Understanding to Preparation
- Lab: From Modeling to Evaluation
- Quiz: From Understanding to Preparation
- Quiz: From Modeling to Evaluation

Module 3: From Deployment to Feedback

- Deployment
- Feedback
- Quiz: From Deployment to Feedback
- Peer-review Assignment

Welcome to Data Science Methodology

Welcome to Data Science Methodology 101!

This is the beginning of a story -one that you'll be telling others about for years to come.

It won't be in the form you experience here, but rather through the stories you'll be sharing with others, as you explain how your understanding of a question resulted in an answer that changed the way something was done.

Despite the recent increase in computing power and access to data over the last couple of decades, our ability to use the data within the decision making process is either lost or not maximized as all too often, we don't have a solid understanding of the questions being asked and how to apply the data correctly to the problem at hand.

Here is a definition of the word methodology.

A methodology is a defined way of

Meth.od.ol.o.gy

- *A system of methods used in a particular area of study of activity.*
    - *"a methodology for investing the concept of focal points"*

It's important to consider it because all too often there is a temptation to bypass methodology and jump directly to solutions. Doing so, however, hinders our best intentions in trying to solve a problem.

This course has one purpose, and that is to share a methodology that can be used within data science, to ensure that the data used in problem solving is relevant and properly manipulated to address the question at hand.

The data science methodology discussed in this course has been outlined by John Rollins, a seasoned and senior data scientist currently practicing at IBM. This course is built on his experience and expresses his position on the importance of following a methodology to be successful.

In a nutshell, the Data Science Methodology aims to answer 10 basic questions in a prescribed sequence.

From problem to approach

1. *What is the problem that you are trying to solve?*
2. *How can you use data to answer the question?*

Working with the data

1. *What data do you need to answer the question?*
2. *Where is the data coming from (identify all sources) and how will you get it?*
3. *Is the data that you collected representative of the problem to be solved?*
4. *What additional work is required to manipulate and work with the data?*

Deriving the answer

1. *In what way can the data be visualized to get the answer that is required?*
2. *Does the model used really answer the initial question or does it need to be adjusted?*
3. *Can you put the model into practice?*
4. *Can you get constructive feedback into answering the question?*

As you can see from this slide, there are two questions designed to define the issue and thus determine the approach to be used; then there are four questions that will help you get organized around the data you will need, and finally there are four additional questions aimed at validating both the data and the approach that gets designed.

Please take a moment now to familiarize yourself with the ten questions, as they will be vital to your success.

This course is comprised of several components:

There are five modules, each going through two stages of the methodology, explaining the rationale as to why each stage is required. Within the same module, a case study is shared that supports what you have just learned. There's also a hands-on lab, which helps to apply the material.

The case study included in the course, highlights how the data science methodology can be applied in context.

It revolves around the following scenario: There is a limited budget for providing healthcare to the public. Hospital readmissions for re-occurring problems can be seen as a sign of failure in the system 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? As you'll see, if the new data science pilot program is successful, it will deliver better patient care by giving physicians new tools to incorporate timely, data-driven information into patient care decisions.

The case study sections display these icons at the top right-hand corner of your screen to help you differentiate theory from practice within each module.

A glossary of data science terms is also provided to assist with clarifying key terms used within the course.

While participating in this course, if you come across challenges, or have questions, then please explore the discussion and wiki sessions.

So, now that you're all set, adjust your headphones and let's get started!

Data Science Methodologies

This course focuses on the Foundational Methodology for Data Science by John Rollins, which was introduced in the previous video. However, it is not the only methodology that you will encounter in data science. For example, in data mining, the Cross Industry Process for Data Mining (CRISP-DM) methodology is widely used.

What is CRISP-DM?

The CRISP-DM methodology is a process aimed at increasing the use of data mining over a wide variety of business applications and industries. The intent is to take case specific scenarios and general behaviors to make them domain neutral. CRISP-DM is comprised of six steps with an entity that has to implement in order to have a reasonable chance of success. The six steps are shown in the following diagram:

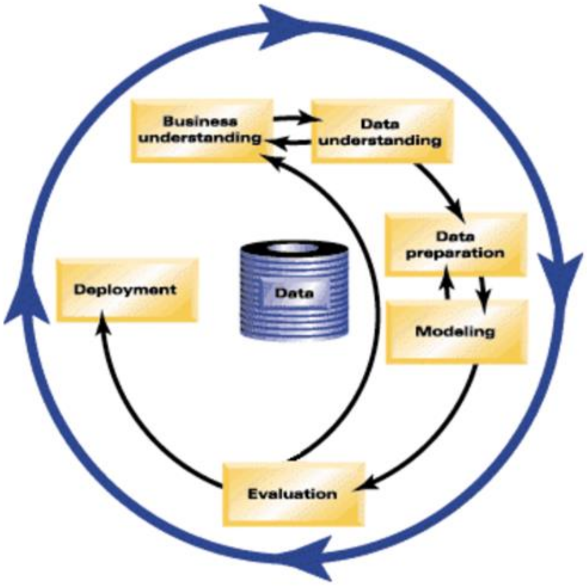


Fig.1 CRISP-DM model, IBM Knowledge Center, CRISP-DM Help Overview

1. **Business Understanding** This stage is the most important because this is where the intention of the project is outlined. Foundational Methodology and CRISP-DM are aligned here. It requires communication and clarity. The difficulty here is that stakeholders have different objectives, biases, and modalities of relating information. They don't all see the same things or in the same manner. Without clear, concise, and complete perspective of what the project goals are resources will be needlessly expended.

2. **Data Understanding** Data understanding relies on business understanding. Data is collected at this stage of the process. The understanding of what the business wants and needs will determine what data is collected, from what sources, and by what methods. CRISP-DM combines the stages of Data Requirements, Data Collection, and Data Understanding from the Foundational Methodology outline.

3. **Data Preparation** Once the data has been collected, it must be transformed into a useable subset unless it is determined that more data is needed. Once a dataset is chosen, it must then be checked for questionable, missing, or ambiguous cases. Data Preparation is common to CRISP-DM and Foundational Methodology.
4. **Modeling** Once prepared for use, the data must be expressed through whatever appropriate models, give meaningful insights, and hopefully new knowledge. This is the purpose of data mining: to create knowledge information that has meaning and utility. The use of models reveals patterns and structures within the data that provide insight into the features of interest. Models are selected on a portion of the data and adjustments are made if necessary. Model selection is an art and science. Both Foundational Methodology and CRISP-DM are required for the subsequent stage.
5. **Evaluation** The selected model must be tested. This is usually done by having a pre-selected test, set to run the trained model on. This will allow you to see the effectiveness of the model on a set it sees as new. Results from this are used to determine efficacy of the model and foreshadows its role in the next and final stage.
6. **Deployment** In the deployment step, the model is used on new data outside of the scope of the dataset and by new stakeholders. The new interactions at this phase might reveal the new variables and needs for the dataset and model. These new challenges could initiate revision of either business needs and actions, or the model and data, or both.

CRISP-DM is a highly flexible and cyclical model. Flexibility is required at each step along with communication to keep the project on track. At any of the six stages, it may be necessary to revisit an earlier stage and make changes. The key point of this process is that it’s cyclical; therefore, even at the finish you are having another business understanding encounter to discuss the viability after deployment. The journey continues.

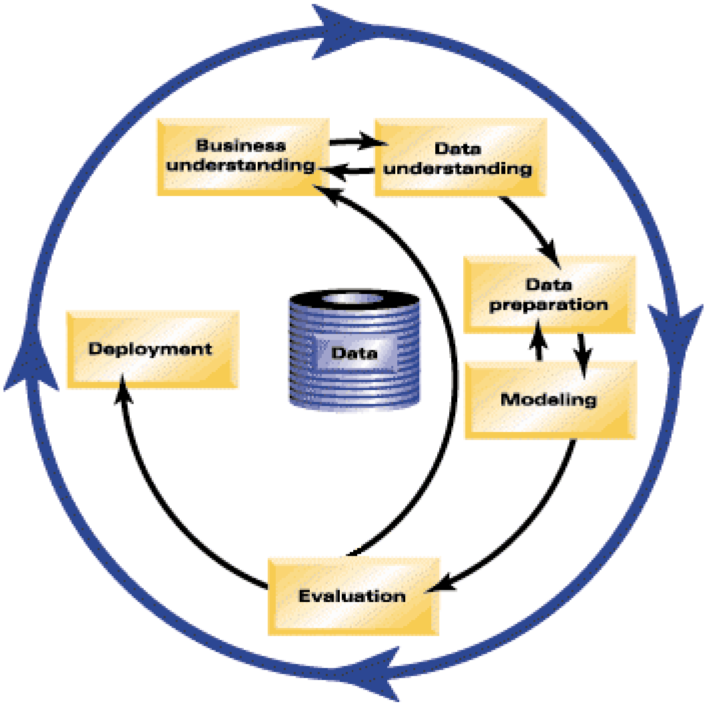
For more information on CRISP-DM, go to: IBM Knowledge Center – CRISP-DM Help Overview

CRISP-DM

CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, is an industry-proven way to guide your data mining efforts.

• As a **methodology**, it includes descriptions of the typical phases of a project, the tasks involved with each phase, and an explanation of the relationships between these tasks.

• As a **process model**, CRISP-DM provides an overview of the data mining life cycle.



The life cycle model consists of six phases with arrows indicating the most important and frequent dependencies between phases. The sequence of the phases is not strict. In fact, most projects move back and forth between phases as necessary.

The CRISP-DM model is flexible and can be customized easily. For example, if your organization aims to detect money laundering, it is likely that you will sift through large amounts of data without a specific modeling goal. Instead of modeling, your work will focus on data exploration and visualization to uncover suspicious patterns in financial data. CRISP-DM allows you to create a data mining model that fits your particular needs.

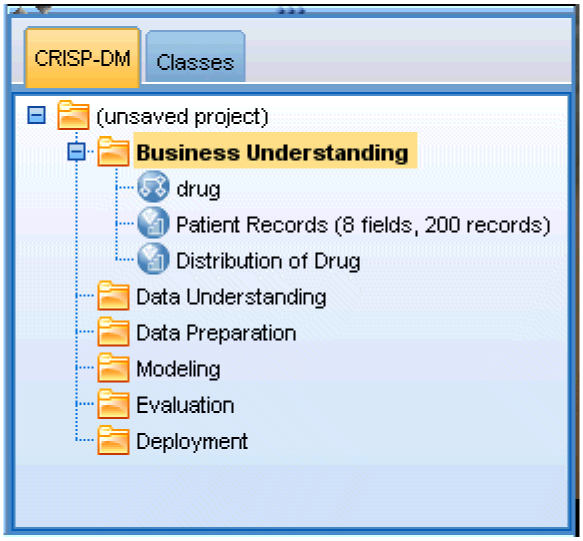
In such a situation, the modeling, evaluation, and deployment phases might be less relevant than the data understanding and preparation phases. However, it is still important to consider some of the questions raised during these later phases for long-term planning and future data mining goals.

IBM® SPSS® Modeler incorporates the CRISP-DM methodology in two ways to provide unique support for effective data mining.

- The CRISP-DM project tool helps you organize project streams, output, and annotations according to the phases of a typical data mining project. You can produce reports at any time during the project based on the notes for streams and CRISP-DM phases.
- Help for CRISP-DM guides you through the process of conducting a data mining project. The help system includes tasks lists for each step as well as examples of how CRISP-DM works in the real world. You can access CRISP-DM Help by choosing CRISP-DM Help from the main window Help menu.

The CRISP-DM project tool provides a structured approach to data mining that can help ensure your project's success. It is essentially an extension of the standard IBM® SPSS® Modeler project tool. In fact, you can toggle between the CRISP-DM view and the standard Classes view to see your streams and output organized by type or by phases of CRISP-DM.

CRISP-DM project tool



CRISP-DM | Classes

- (unsaved project)
  - Business Understanding
    - drug
    - Patient Records (8 fields, 200 records)
    - Distribution of Drug
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment

Using the CRISP-DM view of the project tool, you can:

- Organize a project's streams and output according to data mining phases.
- Take notes on your organization's goals for each phase.
- Create custom tooltips for each phase.
- Take notes on the conclusions drawn from a particular graph or model.
- Generate an HTML report or update for distribution to the project team.

IBM® SPSS® Modeler offers an online guide for the non-proprietary CRISP-DM process model. The guide is organized by project phases and provides the following support:

- An overview and task list for each phase of CRISP-DM
- Help on producing reports for various milestones

- Real-world examples illustrating how a project team can use CRISP-DM to light the way for data mining
- Links to additional resources on CRISP-DM

You can access CRISP-DM Help by choosing CRISP-DM Help from the main window Help menu.

In addition to IBM® SPSS® Modeler support for CRISP-DM, there are several ways to expand your understanding of data mining processes.

- Visit the CRISP-DM consortium Web site at www.crisp-dm.org
- Read the CRISP-DM manual, created by the CRISP-DM consortium and supplied with this release.
- Read Data Mining with Confidence, copyright 2002 by SPSS Inc., ISBN 1-56827-287-1.

Even before working in IBM® SPSS® Modeler, you should take the time to explore what your organization expects to gain from data mining. Try to involve as many key people as possible in these discussions and document the results. The final step of this CRISP-DM phase discusses how to produce a project plan using the information gathered here.

Although this research may seem dispensable, it's not. Getting to know the business reasons for your data mining effort helps to ensure that everyone is on the same page before expending valuable resources.

Click through the steps listed below to get started.

From Problem to Approach

Business Understanding

Welcome to Data Science Methodology 101 From Problem to Approach Business Understanding!

Has this ever happened to you? You've been called into a meeting by your boss, who makes you aware of an important task one with a very tight deadline that absolutely has to be met. You both go back and forth to ensure that all aspects of the task have been considered and the meeting ends with both of you confident that things are on track. Later that afternoon, however, after you've spent some time examining the various issues at play, you realize that you need to ask several additional questions in order to truly accomplish the task. Unfortunately, the boss won't be available again until tomorrow morning. Now, with the tight deadline still ringing in your ears, you start feeling a sense of uneasiness.

So, what do you do?

Do you risk moving forward or do you stop and seek clarification.

Data science methodology begins with spending the time to seek clarification, to attain what can be referred to as a business understanding.

Having this understanding is placed at the beginning of the methodology because getting clarity around the problem to be solved, allows you to determine which data will be used to answer the core question.

Rollins suggests that having a clearly defined question is vital because it ultimately directs the analytic approach that will be needed to address the question.

All too often, much effort is put into answering what people THINK is the question, and while the methods used to address that question might be sound, they don't help to solve the actual problem.

Establishing a clearly defined question starts with understanding the GOAL of the person who is asking the question.

For example, if a business owner asks: "How can we reduce the costs of performing an activity?"

We need to understand, is the goal to improve the efficiency of the activity? Or is it to increase the businesses profitability?

Once the goal is clarified, the next piece of the puzzle is to figure out the objectives that are in support of the goal.

By breaking down the objectives, structured discussions can take place where priorities can be identified in a way that can lead to organizing and planning on how to tackle the problem.

Depending on the problem, different stakeholders will need to be engaged in the discussion to help determine requirements and clarify questions.

So now, let's look at the case study related to applying "Business Understanding" In the case study, the question being asked is: What is the best way to allocate the limited healthcare budget to maximize its use in providing quality care?

This question is one that became a hot topic for an American healthcare insurance provider.

As public funding for readmissions was decreasing, this insurance company was at risk of having to make up for the cost difference, which could potentially increase rates for its customers.

Knowing that raising insurance rates was not going to be a popular move, the insurance company sat down with the health care authorities in its region and brought in IBM data scientists to see how data science could be applied to the question at hand.

Before even starting to collect data, the goals and objectives needed to be defined.

After spending time to determine the goals and objectives, the team prioritized "patient readmissions" as an effective area for review. With the goals and objectives in mind, it was found that approximately 30% of individuals who finish rehab treatment would be readmitted to a rehab center within one year; and that 50% would be readmitted within five years.

After reviewing some records, it was discovered that the patients with congestive heart failure were at the top of the readmission list.

It was further determined that a decision-tree model could be applied to review this scenario, to determine why this was occurring.

To gain the business understanding that would guide the analytics team in formulating and performing their first project, the IBM Data scientists, proposed and delivered an on-site workshop to kick things off.

The key business sponsors involvement throughout the project was critical, in that the sponsor:

- Set overall direction
- Remained engaged and provided guidance.
- Ensured necessary support, where needed.

Finally, four business requirements were identified for whatever model would be built.

Namely:

- Predicting readmission outcomes for those patients with Congestive Heart Failure.
- Predicting readmission risk.

- Understanding the combination of events that led to the predicted outcome
- Applying an easy-to-understand process to new patients, regarding their readmission risk.

This ends the Business Understanding section of this course.

Thanks for watching!

Analytic Approach

Welcome to Data Science Methodology 101 From problem to approach Analytic Approach! Selecting the right analytic approach depends on the question being asked.

The approach involves seeking clarification from the person who is asking the question, so as to be able to pick the most appropriate path or approach.

In this video we'll see how the second stage of the data science methodology is applied.

Once the problem to be addressed is defined, the appropriate analytic approach for the problem is selected in the context of the business requirements.

This is the second stage of the data science methodology.

Once a strong understanding of the question is established, the analytic approach can be selected. This means identifying what type of patterns will be needed to address the question most effectively.

If the question is to determine probabilities of an action, then a predictive model might be used.

If the question is to show relationships, a descriptive approach maybe be required.

This would be one that would look at clusters of similar activities based on events and preferences.

Statistical analysis applies to problems that require counts.

For example, if the question requires a yes/ no answer, then a classification approach to predicting a response would be suitable.

Machine Learning is a field of study that gives computers the ability to learn without being explicitly programmed.

Machine Learning can be used to identify relationships and trends in data that might otherwise not be accessible or identified.

In the case where the question is to learn about human behavior, then an appropriate response would be to use Clustering Association approaches.

So now, let's look at the case study related to applying Analytic Approach.

For the case study, a decision tree classification model was used to identify the combination of conditions leading to each patient's outcome.

In this approach, examining the variables in each of the nodes along each path to a leaf, led to a respective threshold value.

This means the 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.

From this information, 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.

A decision tree classification model is easy for non-data scientists to understand and apply, to score new patients for their risk of readmission.

Clinicians can readily see what conditions are causing a patient to be scored as high-risk and 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 classification approach was chosen for building the Congestive Heart Failure readmission model.

This ends the Analytic Approach section for this course.

Thanks for watching!

Lab: From Problem to Approach

This notebook will demonstrate how to apply the first two stages of the data science methodology to a data science problem.

This course uses a third-party tool, Lab: From Problem to Approach, to enhance your learning experience. The tool will reference basic information like your name, email, and Coursera ID.

In this lesson, you have learned:

- 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.

Data Requirements

Welcome to Data Science Methodology 101 From Requirements to Collection Data Requirements!

If your goal is to make a spaghetti dinner but you don't have the right ingredients to make this dish, then your success will be compromised.

Think of this section of the data science methodology as cooking with data.

Each step is critical in making the meal.

So, if the problem that needs to be resolved is the recipe, so to speak, and data is an ingredient, then the data scientist needs to identify:

which ingredients are required, how to source or the collect them, how to understand or work with them, and how to prepare the data to meet the desired outcome.

Building on the understanding of the problem at hand, and then using the analytical approach selected, the Data Scientist is ready to get started.

Now let's look at some examples of the data requirements within the data science methodology.

Prior to undertaking the data collection and data preparation stages of the methodology, it's vital to define the data requirements for decision-tree classification.

This includes identifying the necessary data content, formats and sources for initial data collection.

So now, let's look at the case study related to applying "Data Requirements".

In the case study, the first task was to define the data requirements for the decision tree classification approach that was selected.

This included selecting a suitable patient cohort from the health insurance providers member base.

In order to compile the complete clinical histories, three criteria were identified for inclusion in the cohort.

First, a patient needed to be admitted as in-patient within the provider service area, so they'd have access to the necessary information.

Second, they focused on patients with a primary diagnosis of congestive heart failure during one full year.

Third, a patient must have had continuous enrollment for at least six months, prior to the primary admission for congestive heart failure, so that complete medical history could be compiled.

Congestive heart failure patients who also had been diagnosed as having other significant medical conditions, were excluded from the cohort because those conditions would cause higher-than-average re-admission rates and, thus, could skew the results.

Then the content, format, and representations of the data needed for decision tree classification were defined.

This modeling technique requires one record per patient, with columns representing the variables in the model.

To model the readmission outcome, there needed to be data covering all aspects of the patient's clinical history.

This content would include admissions, primary, secondary, and tertiary diagnoses, procedures, prescriptions, and other services provided either during hospitalization or throughout patient/doctor visits.

Thus, a particular patient could have thousands of records, representing all their related attributes.

To get to the one record per patient format, the data scientists rolled up the transactional records to the patient level, creating a number of new variables to represent that information.

This was a job for the data preparation stage, so thinking ahead and anticipating subsequent stages is important.

This ends the Data Requirements section for this course.

Thanks for watching!

Data Collection

Welcome to Data Science Methodology 101 From Requirements to Collection Data Collection!

After the initial data collection is performed, an assessment by the data scientist takes place to determine whether or not they have what they need.

As is the case when shopping for ingredients to make a meal, some ingredients might be out of season and more difficult to obtain or cost more than initially thought.

In this phase the data requirements are revised and decisions are made as to whether or not the collection requires more or less data.

Once the data ingredients are collected, then in the data collection stage, the data scientist will have a good understanding of what they will be working with.

Techniques such as descriptive statistics and visualization can be applied to the data set, to assess the content, quality, and initial insights about the data.

Gaps in data will be identified and plans to either fill or make substitutions will have to be made.

In essence, the ingredients are now sitting on the cutting board.

Now let's look at some examples of the data collection stage within the data science methodology.

This stage is undertaken as a follow-up to the data requirements stage.

So now, let's look at the case study related to applying "Data Collection".

Collecting data requires that you know the source or, know where to find the data elements that are needed.

In the context of our case study, these can include: demographic, clinical and coverage information of patients, provider information, claims records, as well as pharmaceutical and other information related to all the diagnoses of the congestive heart failure patients.

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.

For example, this can even be done after getting some intermediate results from the predictive modeling.

If those results suggest that the drug information might be important in obtaining a good model, then the time to try to get it would be invested.

As it turned out though, they were able to build a reasonably good model without this drug information.

DBAs and programmers often work together to extract data from various sources, and then merge it.

This allows for removing redundant data, making it available for the next stage of the methodology, which is data understanding.

At this stage, if necessary, data scientists and analytics team members can discuss various ways to better manage their data, including automating certain processes in the database, so that data collection is easier and faster.

Thanks for watching!

From Requirements to Collection

This notebook will demonstrate how to apply the data requirements and data collection stages of the data science methodology to a data science problem.

This course uses a third-party tool, From Requirements to Collection, to enhance your learning experience. The tool will reference basic information like your name, email, and Coursera ID.

Data Understanding

Welcome to Data Science Methodology 101 From Understanding to Preparation Data Understanding!

Data understanding encompasses all activities related to constructing the data set.

Essentially, the data understanding section of the data science methodology answers the question: Is the data that you collected representative of the problem to be solved?

Let's apply the data understanding stage of our methodology, to the case study we've been examining.

In order to understand the data related to congestive heart failure admissions, descriptive statistics needed to be run against the data columns that would become variables in the model.

First, these statistics included Hearst, 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.

For example, for a categorical variable that has too many distinct values to be informative in a model, the histogram would help them decide how to consolidate those values.

The univariates, statistics, and histograms are also used to assess data quality.

From the information provided, certain values can be re-coded or perhaps even dropped if necessary, such as when a certain variable has many missing values.

The question then becomes, does "missing" mean anything?

Sometimes a missing value might mean "no", or "0" (zero), or at other times it simply means "we don't know". Or, if a variable contains invalid or misleading values, such as a numeric variable called "age" that contains 0 to 100 and also 999, where that "triple-9" actually means "missing", but would be treated as a valid value unless we corrected it.

Initially, the meaning of congestive heart failure admission was decided on the basis of a primary diagnosis of congestive heart failure.

But working through the data understanding stage revealed that the initial definition was not capturing all of the congestive heart failure admissions that were expected, based on clinical experience.

This meant looping back to the data collection stage and adding secondary and tertiary diagnoses, and building a more comprehensive definition of congestive heart failure admission.

This is just one example of the interactive processes in the methodology.

The more one works with the problem and the data, the more one learns and therefore the more refinement that can be done within the model, ultimately leading to a better solution to the problem.

This ends the Data Understanding section of this course.

Thanks for watching!

Data Preparation - Concepts

Welcome to Data Science Methodology 101 From Understanding to Preparation Data Preparation - Concepts!

In a sense, data preparation is similar to washing freshly picked vegetables in so far 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 seventy percent and even up to even ninety percent 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 percent.

This time savings translates into increased time for data scientists to focus on creating models.

To continue with our cooking metaphor, we know that the process of chopping onions to a finer state will allow for its flavors to spread through a sauce more easily than that would be the case if we were to drop the whole onion into the sauce pot.

Similarly, 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.

Specifically, the data preparation stage of the methodology answers the question: What are the ways in which data is prepared?

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.

Feature engineering is also part of data preparation.

It is the process of using domain knowledge of the data to create features that make the machine learning algorithms work.

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.

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 detail in this area.

After all, it takes just one bad ingredient to ruin a fine meal.

This ends the Data Preparation section of this course, in which we've reviewed key concepts.

Thanks for watching!

Please note that the phrase "literary review" in the next video: Data Preparation - Case Study, is supposed to be "literature review"

Data Preparation - Case Study

Welcome to Data Science Methodology 101 From Understanding to Preparation Data Preparation - Case Study!

In a sense, data preparation is similar to washing freshly picked vegetables in so far as unwanted elements, such as dirt or imperfections, are removed.

So now, let's look at the case study related to applying Data Preparation concepts.

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.

Transactional records included professional provider facility 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. A given patient could easily have hundreds or even thousands of these records, depending on their clinical history.

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.

For example, frequency and most recent visits to doctors, clinics and hospitals with diagnoses, procedures, prescriptions, and so forth.

Co-morbidities with congestive heart failure were also considered, such as diabetes, hypertension, and many other diseases and chronic conditions that could impact the risk of re-admission for congestive heart failure.

During discussions around data preparation, a literary review on congestive heart failure was also undertaken to see whether any important data elements were overlooked, such as co-morbidities that had not yet been accounted for.

The literary review involved looping back to the data collection stage to add a few more indicators for conditions and procedures.

Aggregating the transactional data at the patient level, meant 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, 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.

This ends the Data Preparation section of this course, in which we applied the key concepts to the case study.

Thanks for watching!

From Understanding to Preparation

This notebook will demonstrate how to apply the data understanding and data preparation stages of the data science methodology to a data science problem. This course uses a third-party tool, From Understanding to Preparation, to enhance your learning experience. The tool will reference basic information like your name, email, and Coursera ID.

In this lesson, you have learned:

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

Modeling - Concepts

Welcome to Data Science Methodology 101 From Modeling to Evaluation Modeling - Concepts!

Modelling 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! This portion of the course is geared toward answering two key questions:

First, what is the purpose of data modeling, and second, what are some characteristics of this process?

Data Modelling focuses on developing models that are either descriptive or predictive.

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.

The data scientist will use a training set for predictive modelling.

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.

The success of data compilation, preparation and modelling, 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.

In John Rollins' descriptive Data Science Methodology, the framework is geared to do 3 things: ***First,*** understand the question at hand. ***Second,*** select an analytic approach or method to solve the problem, and ***third,*** 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.

With dinner just about to be served and a hungry guest at the table, the key question is: Have I made enough to eat?

Well, let's hope so.

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.

This ends the Modeling to Evaluation section of this course, in which we reviewed the key concepts related to modeling. Thanks for watching!

Modeling - Case Study

Welcome to Data Science Methodology 101 From Modeling to Evaluation Modeling - Case Study!

Modelling 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!

Now, let's apply the case study to the modeling stage within the data science methodology.

Here, we'll discuss one of the many aspects of model building, in this case, parameter tuning to improve the 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.

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.

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.

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 parameter.

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.

This concludes the Modeling section of the course, in which we applied the Case Study to the modeling stage within the data science methodology.

Thanks for watching!

Evaluation

Welcome to Data Science Methodology 101 From Modeling to Evaluation - Evaluation!

A model evaluation goes hand-in-hand with model building as such, the modeling and evaluation stages are done iteratively.

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?

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.

So now, let's go back to our case study so that we can apply the "Evaluation" component within the data science methodology.

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.

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.

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?

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 base line.

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.

This ends the Evaluation section of this course.

Thanks for watching!

From Modeling to Evaluation

This notebook will demonstrate how to apply the modeling and evaluation stages of the data science methodology to a data science problem.

This course uses a third-party tool, From Modeling to Evaluation, to enhance your learning experience. The tool will reference basic information like your name, email, and Coursera ID.

In this lesson, you have learned:

- 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.

Deployment

Welcome to Data Science Methodology 101 From Deployment to Feedback - Deployment!

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.

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.

So now, let's look at the case study related to applying Deployment.

- 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.

- 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.

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.

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.

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.

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.

This ends the Deployment section of this course.

Thanks for watching!

Feedback

Welcome to the Data Science Methodology 101 From Deployment to Feedback - Feedback!

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.

That's the way John Rollins sees it and hopefully you do too.

Once the model is evaluated and the data scientist is confident it'll work, it is deployed and put to the ultimate test: actual, real-time use in the field.

So now, let's look at our case study again, to see how the Feedback portion of the methodology is applied.

The plan for the feedback stage included these steps:

- 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 re-admissions.

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.

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.

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.
- 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.

This is the end of the Feedback portion of this course.

Thanks for watching!

Course Summary

Welcome to Data Science Methodology 101 Course Summary! We've come to the end of our story, one that we hope you'll share.

You've learned how to think like a data scientist, including taking the steps involved in tackling a data science problem and applying them to interesting, real-world examples.

These steps have included:

- forming a concrete business or research problem,
- collecting and analyzing data,
- building a model, and
- understanding the feedback after model deployment.

In this course, you've also learned methodical ways of moving from problem to approach,

- including the importance of understanding the question, the business goals and objectives, and
- picking the most effective analytic approach to answer the question and solve the problem.

You've also learned methodical ways of working with the data, specifically,

- determining the data requirements,
- collecting the appropriate data,
- understanding the data, and then
- preparing the data for modeling!

You've also learned how to model the data by using the appropriate analytic approach, based on the data requirements and the problem that you were trying to solve Once the approach was selected,

- you learned the steps involved in evaluating and deploying the model,
- getting feedback on it, and
- using that feedback constructively so as to improve the model.

Remember that the stages of this methodology are iterative!

This means that the model can always be improved for as long as the solution is needed, regardless of whether the improvements come from constructive feedback, or from examining newly available data sources.

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.

And finally, you learned, in a nutshell, the true meaning of a methodology!

That its purpose is to explain how to look at a problem, work with data in support of solving the problem, and come up with an answer that addresses the root problem.

By answering 10 simple questions methodically, we've taught you that a methodology can help you solve not only your data science problems, but also any other problem.

In a nutshell, the Data Science Methodology aims to answer 10 basic questions in a prescribed sequence.

From problem to approach

3. *What is the problem that you are trying to solve?*
4. *How can you use data to answer the question?*

Working with the data

5. *What data do you need to answer the question?*
6. *Where is the data coming from (identify all sources) and how will you get it?*
7. *Is the data that you collected representative of the problem to be solved?*
8. *What additional work is required to manipulate and work with the data?*

Deriving the answer

5. *In what way can the data be visualized to get the answer that is required?*
6. *Does the model used really answer the initial question or does it need to be adjusted?*
7. *Can you put the model into practice?*
8. *Can you get constructive feedback into answering the question?*

Your success within the data science field depends on your ability to apply the right tools, at the right time, in the right order, to the address the right problem. And that is the way John Rollins sees it!

We hope you've enjoyed taking the Data Science Methodology course and found it to be a valuable experience one that you'll share with others!

Thanks for watching!