In this module, you will learn about why we are interested in data science, what a methodology is, and why data scientists need a methodology. You will also learn about the data science methodology and its flowchart. You will learn about the first two stages of the data science methodology, namely Business Understanding and Analytic Approach. Finally, through a lab session, you will also obtain how to complete the Business Understanding and the Analytic Approach stages and the Data Requirements and Data Collection stages pertaining to any data science problem.

**Learning Objectives**

* List the six key stages of the Cross-Industry Process for Data Mining Methodology (CRISP-DM), an industry-standard data science methodology.
* Analyze the first four phases of CRISP-DM.
* Apply the first four phases of the data science methodology to a case study.
* Write clearly defined questions that address a business problem.
* Analyze a case study to determine data requirements.
* Apply the data science methodology to a case study.
* Determine data content, data formats, and data sources prior to data collection and data preparation phases.
* Create a decision tree to classify outcomes in a case study.
* Identify appropriate data sources to address a business problem.

# **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!

(music)

# **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 behaviour, 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!

(music)

# **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 the 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 to 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!

(music)

# **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!

(music)