Data Science with Python Programming

- Presentation By Uplatz
- Contact us: https://training.uplatz.com
- Email: info@uplatz.com
- Phone: +44 7836 212635

Uplatz

Data Science Methodology (Part-1)

Learning outcomes:

In Data Science methodology you will learn about:

- The major steps involved in tackling a data science problem.
- The major steps involved in practicing data science, from forming a concrete business or research problem, to collecting and analysing data, to building a model, and understanding the feedback after model deployment.
- How data scientists think through tackling interesting real-world examples?

Learning outcomes:

Module 1: From Problem to Approach

Business Understanding

Analytic Approach

Module 2: From Requirements to Collection

Data Requirements

Data Collection

Module 3: From Understanding to Preparation

Data Understanding

Data Preparation

Module 4: From Modeling to Evaluation

Modeling

Evaluation

Module 5: From Deployment to Feedback

Deployment

Feedback

Learning outcomes:

Module 1: From Problem to Approach

- Business Understanding
- Analytic Approach

Module 2: From Requirements to Collection

- Data Requirements
- Data Collection

Module 3: From Understanding to Preparation

- Data Understanding
- Data Preparation

Module 1: From Problem to Approach

Business Understanding:

Welcome to Data Science Methodology 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?

Module 1: From Problem to Approach

Business Understanding:

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.

Module 1: From Problem to Approach

Business Understanding:

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.

Module 1: From Problem to Approach

Business Understanding:

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.

Module 1: From Problem to Approach

Business Understanding:

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 centre 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.

Module 1: From Problem to Approach

Business Understanding:

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.

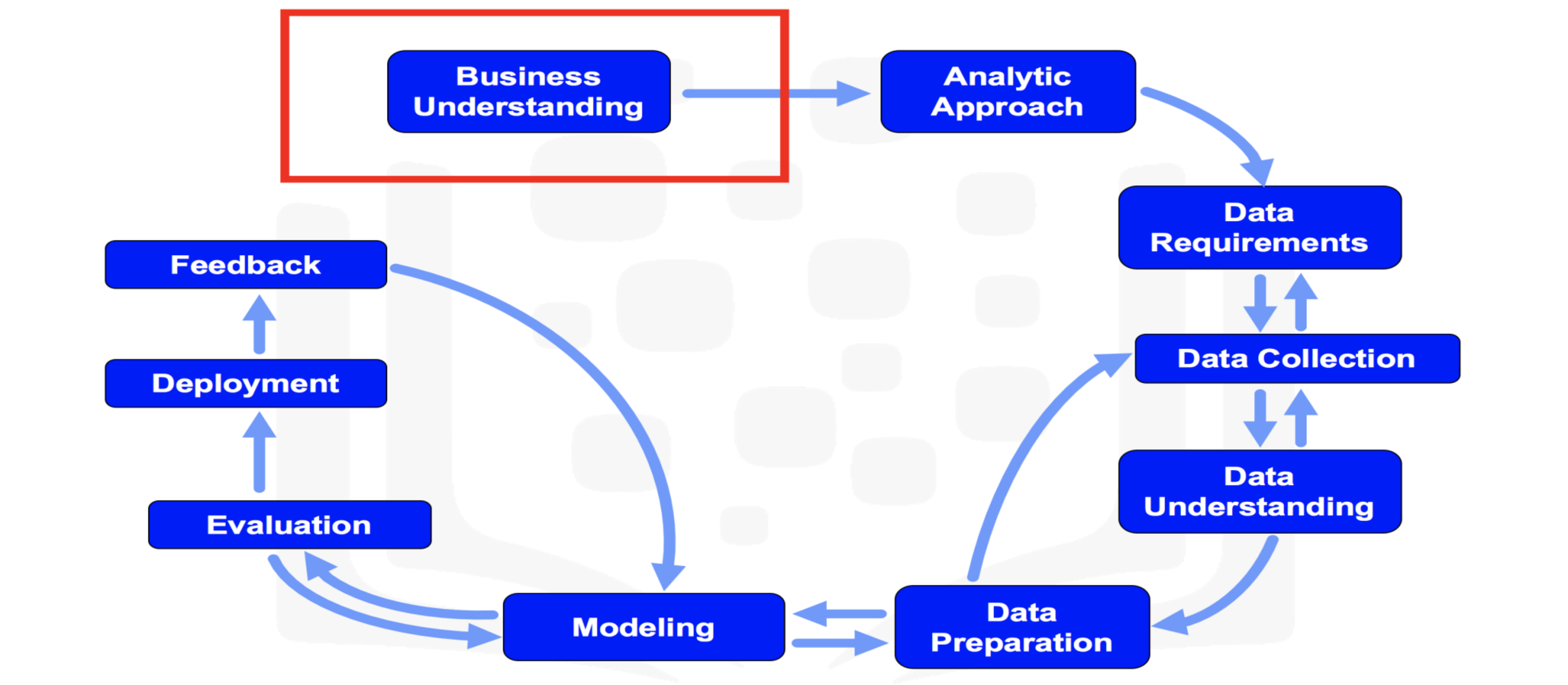
Understanding the combination of events that led to the predicted outcome

Applying an easy-to-understand process to new patients, regarding their readmission risk.

Module 1: From Problem to Approach

Business Understanding:

This is the **Data Science Methodology**, a flowchart that begins with business understanding.



Module 1: From Problem to Approach

Business Understanding:

Looking at this diagram, we immediately spot two outstanding features of the data science methodology.

What are they?

1. The flowchart is highly iterative.
2. The flowchart never ends.

Module 1: 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. 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

Module 1: From Problem to Approach

Analytic Approach:

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.

Module 1: From Problem to Approach

Analytic Approach:

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.

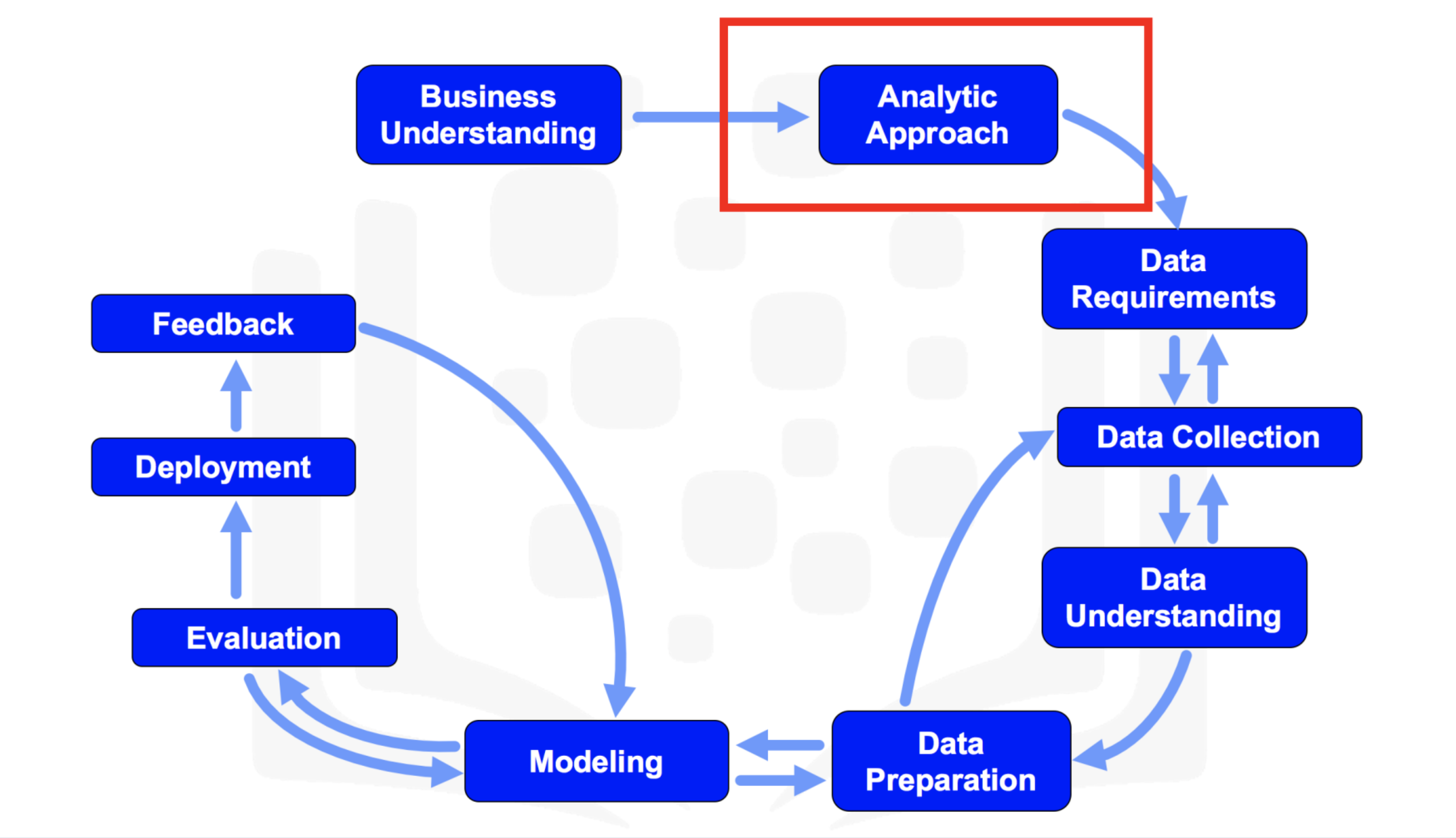
Module 1: From Problem to Approach

Analytic Approach:

From this information, the analysts can obtain the re-admission 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.

Module 1: From Problem to Approach

Analytic Approach:



Uplatz

Module 1: From Problem to Approach

Analytic Approach:

So why are we interested in data science?

Once the business problem has been clearly stated, the data scientist can define the analytic approach to solve the problem. This step entails expressing the problem in the context of statistical and machine-learning techniques, so that the entity or stakeholders with the problem can identify the most suitable techniques for the desired outcome.

Why is the analytic approach stage important?

Because it helps identify what type of patterns will be needed to address the question most effectively.

Module 2: From Requirements to Collection

Data Requirements:

Welcome to Data Science Methodology 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.

Module 2: From Requirements to Collection

Data Requirements:

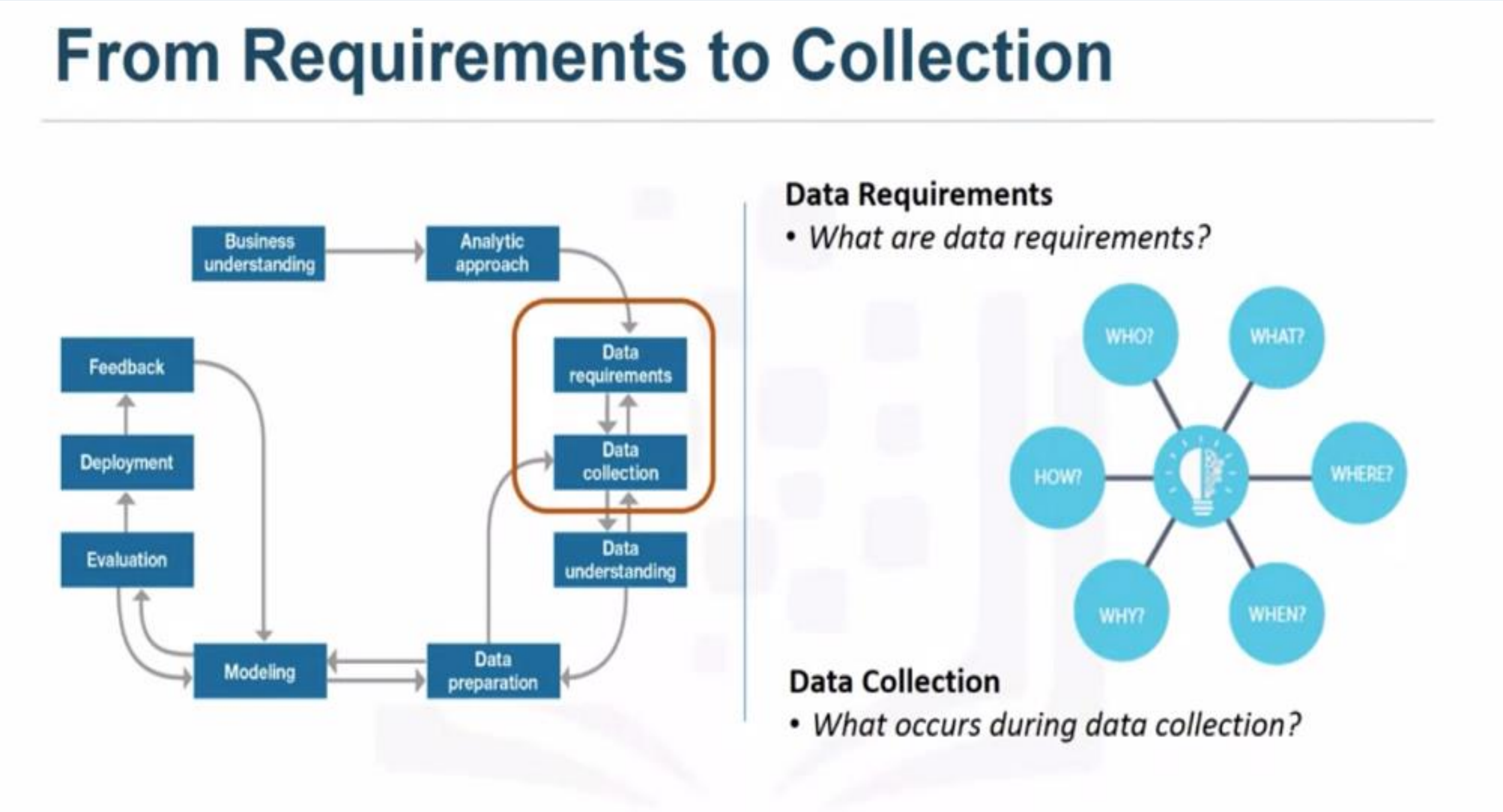
Once we have found a way to solve our problem, we will need to discover the correct data for our model.

Data Requirements is the stage where we identify the necessary data content, formats, and sources for initial data collection, and we use this data inside the algorithm of the approach we chose.

During the process of data requirements, one should find the answers for questions like ‘what’, ‘where’, ‘when’, ‘why’, ‘how’ & ‘who’.

Module 2: From Requirements to Collection

Data Requirements:



From Requirements to Collection

Data Requirements

- *What are data requirements?*

Data Collection

- *What occurs during data collection?*

Module 2: 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.

Module 2: From Requirements to Collection

Data Collection:

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. In the initial data collection stage, data scientists identify and gather the available data resources. These can be in the form of structured, unstructured, and even semi-structured data relevant to the problem domain.

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.

Module 3: 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?

The objectives of data understanding are:

Understand the attributes of the data.

Summarize the data by identifying key characteristics, such as data volume and total number of variables in the data.

Understand the problems with the data, such as missing values, inaccuracies, and outliers.

Visualize the data to validate the key characteristics of the data or unearth problems with the summary statistics.

Module 3: From Understanding to Preparation

Data Understanding:

Visualization of a variable helps in obtaining insights into the distribution of the data. The two common visualizations for single variables are a histogram and a box plot. In data understanding phase one typically

- Understand data touch points in the context of business process
- Gather knowledge on where data originates from, how it gets processed, what decisions are being made, where it is getting stored and how it flows to downstream
- Deep dive into business meaning of the data being leveraged as well as knowledge present in existing system in form of rule
- Check if it will be appropriate to use additional industry known external data sources that can enhance decision boundary
- Check for target label availability as well as check for late arriving labels

Module 3: From Understanding to Preparation

Data Preparation:

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 flavours 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.

Module 3: From Understanding to Preparation

Data Preparation:

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.

Module 3: From Understanding to Preparation

Data Preparation:

Feature engineering is critical when machine learning tools are being applied to analyse 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.

Thank you