

Data Science Methodology: 10 Steps For Best Solutions

In Data Science, this is called **Data Science Methodology** — an iterative process with a prescribed sequence of steps that are followed by data scientists to approach a problem and find a solution. It is a cyclic process that guides business analysts and data scientists to perform suitably.

For example, a company needs to know what features to include in their product or service to make it successful. They approach a business analyst or a data scientist to find a solution. A number of factors can be considered when thinking of the solution.

There is also a need to understand what success means with respect to this certain problem, it could just mean purely creating profits for the business, or it could mean customer satisfaction and their interaction with the product or how their service is affecting the market. In such cases, using the Data Science Methodology has proved to be an efficient and effective method.

Data Science Methodology comprises of ten steps that are repeated constantly for data scientists to arrive at the best solution.

These can be combined into five sections:

From Problem to Approach which includes the Business Understanding and Analytical Approach stages.

From Requirements to Collection under which the Data Requirements and Data collection stages are present.

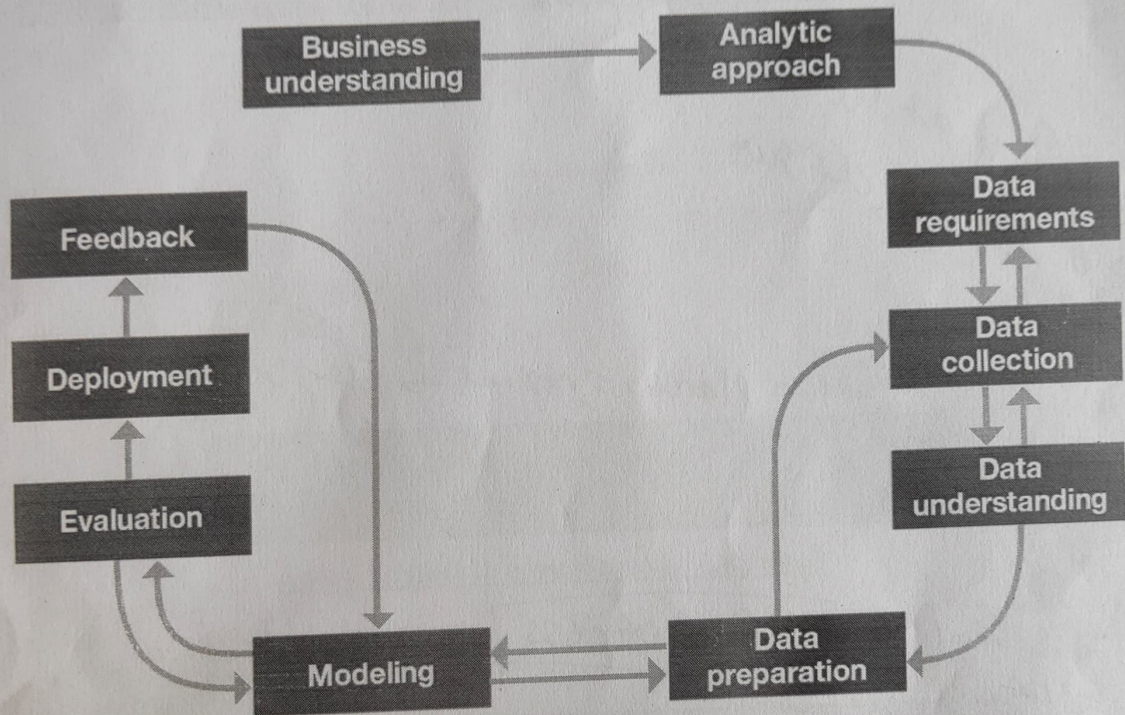
From Understanding to Preparation which involves the Data Understanding and Data Preparation stages.

From Modeling to Evaluation which includes the Modeling and Evaluation stages.

And lastly, **From Deployment to Feedback** under which the Deployment and Feedback stages are included.

What is Methodology in Data Science?

Methodology in Data Science is the best way to organize your work, doing it better, and without losing time. Data Science Methodology is composed of 10 parts:



JIRA
 { Plan
 Track
 manage

2. Login Screen :

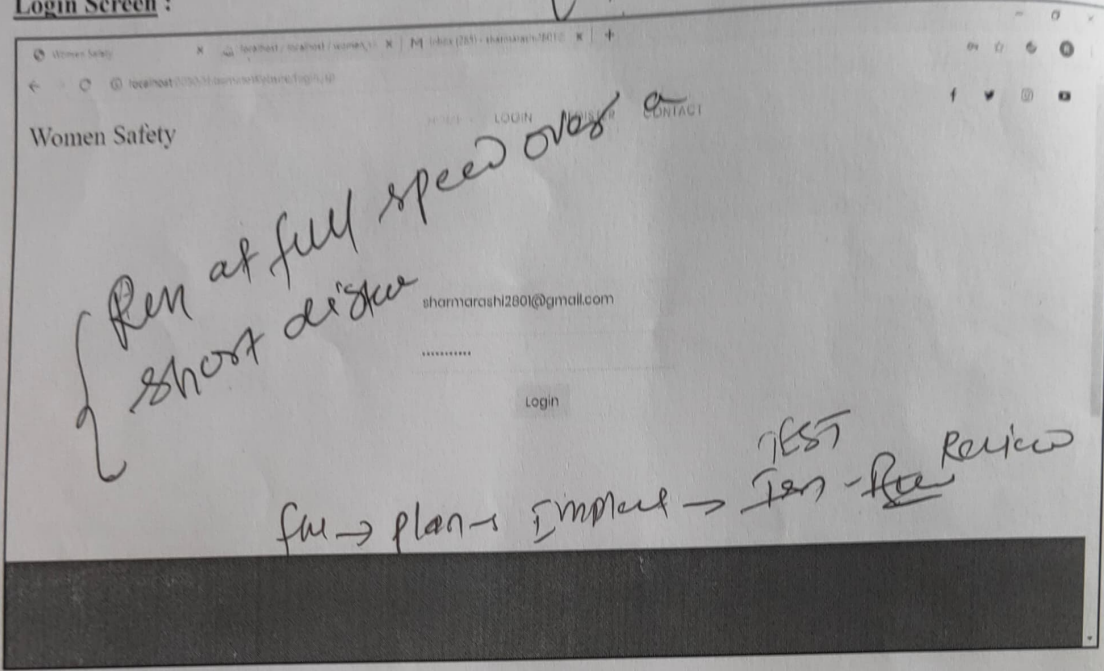


Figure 4.6 : Login Screen of Quick Contact

Web develop.
 S/W
 Data Science

Agile process

S/W process

project -> pre-requisite

Product Manager

Analyst Manager

REQ. gathering

Product Manager

3 months

30 story / task -> 3 month -> 12 week
 1 sprint -> 2 week
 6 sprints -> 12 week

Sprint

Q.A. Sub

SP2

JIRA

QA

QA Acceptance User

QA User

Product Manager

Agile Method

Table of Contents

- 10 Steps of Data Science Methodology
 - 1. Business Understanding
 - 2. Analytic Approach
 - 3. Data Requirements
 - 4. Data Collection
 - 5. Data Understanding
 - 6. Data Preparation
 - 7. Modeling
 - 8. Evaluation
 - 9. Deployment
 - 10. Feedback
- Conclusion
- Where is the analytic approach used in data science?
- What happens in the modeling stage of data science methodology?
- Why are data science and its methodology important?

10 Steps of Data Science Methodology

1. Business Understanding

For any project or problem-solving, the first stage is always understanding the business. This involves defining the problem, project objectives, and requirements of the solutions. This step plays a critical role in defining how the project will develop. A thorough discussion with the clients, understanding how their business works, requirements from the product or service, and clarifying each aspect of the problem can take time and prove to be laborious, but it is a necessity.

2. Analytic Approach

After the problem has been clearly defined, the analytical approach which will be used to solve the problem can be defined. This means expressing the problem in the framework of statistical and machine learning techniques. There are different models that can be used and it depends on the type of outcome needed.

Statistical analysis can be used if it requires summarising, counting, finding trends in the data. To assess the relationships between various elements and the environment and how they affect each other, a descriptive model can be used.

And for predicting the possible outcomes or calculating the probabilities, a predictive model can be used which is a data mining technique. A training set that is a set of historical data that includes its outcomes, is used for predictive modeling.

3. Data Requirements

The analytical approach chosen in the previous stage defines the kind of data needed to solve the problem. This step identifies the data contents, formats, and the sources for data collection. The data selected should be able to answer all the 'what', 'who', 'when', 'where', 'why' and 'how' questions about the problem.

4. Data Collection

In the fourth stage, the data scientist identifies all the data resources and collects data in all forms such as structured, unstructured, and semi-structured data that is relevant to the problem. Data is available on many websites and there are premade datasets that can also be used.

At times, if there is a requirement for important data that is not accessible freely, certain investments need to be made in order to obtain such datasets. If later there are any gaps identified within the collected data that is hindering the project development, the data scientist has to revise the requirements and collect more data.

The more the data acquired, the better the models will be built that can produce more effective outcomes.

5. Data Understanding

In this stage, the data scientist tries to understand the data collected. This involves applying descriptive analysis and visualization techniques to the

data. This will help in a better understanding of the data content and the quality of the data and developing initial insights from the data. If there are any gaps identified in this step, the data scientist can go back to the previous step and gather more data.

6. Data Preparation

This stage comprises all the activities needed to construct the data to make it suitable to be used for the modeling stage. This includes data cleaning i.e. managing missing data, deleting duplicates, changing the data into a uniform format, etc., combining data from various sources, and transforming data into useful variables.

This is one of the most time-consuming steps. However, there are automated methods available today that can accelerate the process of data preparation. At the end of this stage, only the data needed to solve the problem is retained to make the model run smoothly with minimal errors.

7. Modeling

The dataset prepared in the previous stage is used for creating the modeling stage. Here the type of model to be used is defined by the approach decided upon in the analytical approach stage. Thus, the kind of dataset varies depending on whether it is a descriptive, predictive approach or a statistical analysis.

This is one of the most iterative processes in the methodology as the data scientist will use multiple algorithms to arrive at the best model for the chosen variables. It also involves combining various business insights that are continuously being discovered which leads to refining the prepared data and model.

8. Evaluation

The data scientist evaluates the quality of the model and ensures that it meets all the requirements of the business problem. This involves the model undergoing various diagnostic measures and statistical significance testing. It helps in interpreting the efficacy with which the model arrives at a solution.

9. Deployment

Once the model has been developed and approved by the business clients and other stakeholders involved, it is deployed into the market. It could be deployed to a set of users or into a test environment. Initially, it might be

introduced in a limited way, until it is tested completely and been successful in all its aspects.

10. Feedback

The last stage in the methodology is feedback. This includes results collected from the deployment of the model, feedback on the model's performance from the users and clients, and observations from how the model works in the deployed environment.

Data scientists analyze the feedback received, which helps them refine the model. It is also a highly iterative stage as there is a continuous back and forth between the modeling and feedback stages. This process continues till the model is providing satisfactory and acceptable results.