EMBA Program
MB-511

Data Science for Managerial Decisions (MB 511)
Building Data Science Solutions

Instructor
Anant Prakash Awasthi

EMBA Program
MB-511

Building a Data Science Solution

Approaches – Methodologies

EMBA Program
MB-511

Methodologies play a crucial role in building data science solutions by providing structure, guiding decision-making, and ensuring that projects are executed efficiently and effectively.

- Providing a Structured Framework
- Facilitating Collaboration
- Ensuring Alignment with Business Objectives
- Enhancing Data Quality and Preparation
- Improving Model Performance and Evaluation
- Managing Complexity and Risk
- Fostering Iteration and Continuous Improvement
- Guiding Deployment and Operationalization
- Ensuring Accountability and Documentation
- Optimizing Resource Allocation
- Improving Stakeholder Engagement and Communication
- Supporting Scalability and Reproducibility

Data Science Project Execution Methodologies

EMBA Program
MB-511

- **CRISP-DM (Cross-Industry Standard Process for Data Mining)**
- SEMMA (Sample, Explore, Modify, Model, Assess)
- OODA Loop (Observe, Orient, Decide, Act)
- KDD (Knowledge Discovery in Databases)
- Team Data Science Process (TDSP)
- Lean AI/Agile Data Science
- Google's TFX (TensorFlow Extended) Pipeline for Machine Learning
- ASUM-DM (Analytics Solutions Unified Method for Data Mining)
- Design Thinking for Data Science
- End-to-End Machine Learning Pipeline

CRISP-DM (Cross-Industry Standard Process for Data Mining)

EMBA Program
MB-511

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a widely used, industry-standard methodology for data science projects. Developed in the late 1990s, it provides a structured approach to building data-driven solutions, particularly in the context of data mining and predictive modeling. CRISP-DM is highly adaptable and is applied across a variety of industries, from finance to healthcare, retail, and more.

Why CRISP-DM:

- Industry-Agnostic
- Iterative and Non-Linear Process
- Emphasizes Business Understanding
- Data-Centric
- Model Agnostic
- Evaluation-Focused
- Deployment and Monitoring
- Well-Established and Widely Used
- Phases Are Well-Defined and Modular
- Supports Both Supervised and Unsupervised Learning
- Encourages Documentation and Transparency
- Focuses on Practical Implementation

CRISP-DM (Cross-Industry Standard Process for Data Mining)

Characteristics

EMBA Program
MB-511

Industry-Agnostic

- CRISP-DM is designed to be adaptable to any industry, from finance to healthcare, retail, and beyond. It provides a general framework that can be applied to various types of data mining and machine learning problems without being tied to a specific domain.
- Benefit: It can be used in diverse fields and is highly versatile, allowing for broad application.

Iterative and Non-Linear Process

- Although CRISP-DM outlines a structured sequence of six phases, the process is iterative rather than strictly linear. Data scientists often need to loop back to previous stages based on new insights or issues that arise during later stages.
- Benefit: Encourages continuous improvement and refinement of models and approaches, ensuring that the solution evolves with the project.

CRISP-DM (Cross-Industry Standard Process for Data Mining)

Characteristics

EMBA Program
MB-511

Emphasizes Business Understanding

- CRISP-DM places significant importance on understanding the business problem before diving into data. The very first phase, "Business Understanding," ensures that the project aligns with real-world business objectives and that the solution delivers practical value.
- Benefit: Keeps the focus on solving business problems, ensuring the project's outcome is useful and relevant to stakeholders.

Data-Centric

- The methodology emphasizes the importance of data at every step. It has dedicated phases for understanding, preparing, and exploring data, which ensures that data is treated as a crucial asset.
- Benefit: Forces data scientists to thoroughly explore and prepare data, leading to higher-quality models and insights.

CRISP-DM (Cross-Industry Standard Process for Data Mining)

Characteristics

EMBA Program
MB-511

Model Agnostic

- CRISP-DM does not prescribe specific modeling techniques or algorithms. It allows for the use of any method that fits the problem at hand, whether it is regression, decision trees, clustering, or deep learning.
- Benefit: Provides flexibility in choosing the best modeling approach for the given problem.

Evaluation-Focused

- Before moving forward to deployment, CRISP-DM emphasizes rigorous evaluation of the model's performance against the business objectives. This ensures that the solution not only works from a technical perspective but also meets the predefined success criteria.
- Benefit: Ensures that only well-performing models are deployed, reducing the risk of failure in production environments.

CRISP-DM (Cross-Industry Standard Process for Data Mining)

Characteristics

EMBA Program
MB-511

Deployment and Monitoring

The final phase of CRISP-DM includes deploying the model into production and monitoring its performance over time. This phase ensures that the solution is fully operational and integrated into the organization's workflows.

Benefit: Helps transition from a prototype model to a fully functional, real-world solution that delivers continuous value.

Well-Established and Widely Used

CRISP-DM is one of the most established and commonly used methodologies in the field of data science. Its widespread adoption means that many professionals are familiar with it, making it easier for teams to collaborate and follow a shared process.

Benefit: Reduced learning curve and increased compatibility across teams and projects.

CRISP-DM (Cross-Industry Standard Process for Data Mining)

Characteristics

EMBA Program
MB-511

Phases Are Well-Defined and Modular

Each of the six phases in CRISP-DM—Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment—is clearly defined, which makes it easy to know what tasks to focus on at any given point.

Benefit: Provides clarity and structure, making it easy to plan and manage projects effectively.

Supports Both Supervised and Unsupervised Learning

CRISP-DM can be applied to a wide range of data science tasks, whether the goal is to predict outcomes (supervised learning) or find patterns in data (unsupervised learning).

Benefit: Versatile for various types of projects, from predictive modeling to clustering and segmentation.

CRISP-DM (Cross-Industry Standard Process for Data Mining)

Characteristics

EMBA Program
MB-511

Encourages Documentation and Transparency

CRISP-DM encourages thorough documentation at each stage, ensuring that the project is well-documented and that key decisions are recorded. This helps ensure transparency and traceability, especially when models need to be retrained or audited.

Benefit: Facilitates knowledge sharing, reproducibility, and accountability within data science teams.

Focuses on Practical Implementation

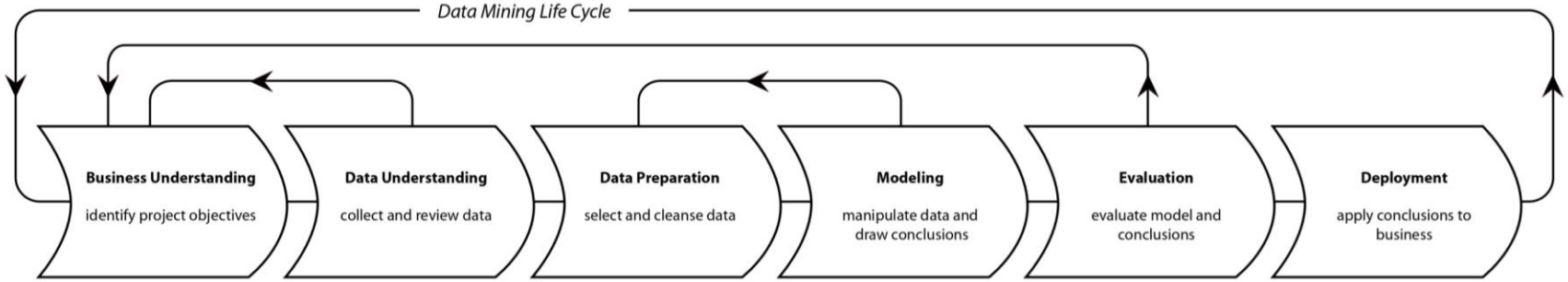
CRISP-DM is not just about theoretical models or academic research. The final phase—deployment—ensures that models are implemented and deliver tangible results for the business.

Benefit: Encourages real-world application and focuses on delivering value rather than just building models.

CRISP-DM (Cross-Industry Standard Process for Data Mining)

Workflow

EMBA Program
MB-511



Determine Business Objectives
Background
Business Objectives
Business Success Criteria
(Log and Report Process)

Assess Situation
Inventory of Resources, Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits
(Log and Report Process)

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria
(Log and Report Process)

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques
(Log and Report Process)

Collect Initial Data
Initial Data Collection Report
(Log and Report Process)

Describe Data
Data Description Report
(Log and Report Process)

Explore Data
Data Exploration Report
(Log and Report Process)

Verify Data Quality
Data Quality Report
(Log and Report Process)

Data Set
Data Set Description
(Log and Report Process)

Select Data
Rationale for Inclusion/ Exclusion
(Log and Report Process)

Clean Data
Data Cleaning Report
(Log and Report Process)

Construct Data
Derived Attributes
Generated Records
(Log and Report Process)

Integrate Data
Merged Data
(Log and Report Process)

Format Data
Reformatted Data
(Log and Report Process)

Select Modeling Technique
Modeling Technique
Modeling Assumptions
(Log and Report Process)

Generate Test Design
Test Design
(Log and Report Process)

Build Model Parameter Settings
Models
Model Description
(Log and Report Process)

Assess Model
Model Assessment
Revised Parameter
(Log and Report Process)

Evaluate Results
Align Assessment of Data Mining Results with Business Success Criteria
(Log and Report Process)

Approved Models
Review Process
Review of Process
(Log and Report Process)

Determine Next Steps
List of Possible Actions
Decision
(Log and Report Process)

Plan Deployment
Deployment Plan
(Log and Report Process)

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan
(Log and Report Process)

Produce Final Report
Final Report
Final Presentation
(Log and Report Process)

Review Project
Experience Documentation
(Log and Report Process)

Generic Tasks
Specialized Tasks
(Process Instances)

a visual guide to CRISP-DM methodology

SOURCE CRISP-DM 1.0
http://www.crisp-dm.org/download.htm
DESIGN Nicole Leaper
http://www.nicoleleaper.com

EMBA Program
MB-511

Have a question?

Feel Free to Reach out at

- +91-88846-92929 (WhatsApp)
- anant.awasthi@outlook.com (E-Mail)