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

> Разделы: Бизнес-задачи
> Loginom: Data Mining

Открытый стандарт моделирования процессов, описывающий общие подходы, используемые в интеллектуальном анализе данных.

Стандарт CRISP-DM начал разрабатываться 1996 году консорциумом пяти компаний, которые имели значительный собственный опыт использования аналитических технологий: Integral Solutions, Teradata, Daimler AG, NCR Corporation и OHRA.

Первая версия методологии была представлена на 4-м семинаре CRISP-DM SIG в Брюсселе в марте 1999 года и опубликована в виде пошагового руководства по интеллектуальному анализу данных в конце того же года. Между 2006 и 2008 годами был выпущен CRISP-DM 2.0 и обсуждались вопросы обновления модели процесса.

CRISP-DM стал наиболее широко используемой формой модели интеллектуального анализа данных. Недостатком этой модели является отсутствие средств поддержки управления проектами. Преимущество CRISP-DM заключается в том, что он нейтрален в отношении предметных областей, инструментов и приложений.

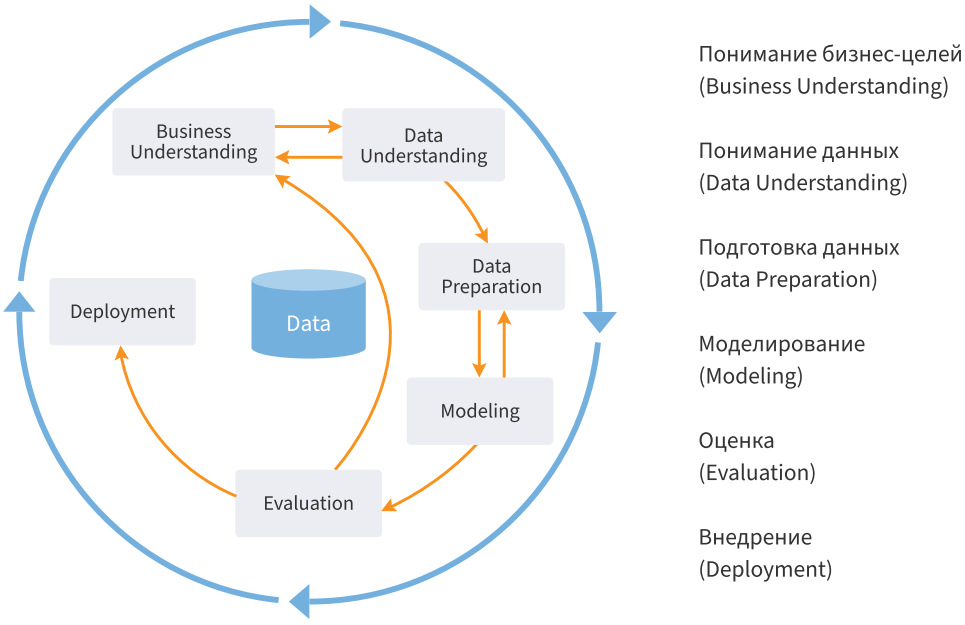
Стандарт CRISP-DM представляет процесс интеллектуального анализ данных в виде шести фаз:

- **Понимание бизнес-целей** (Business Understanding). На данном этапе производится исследование бизнес-процессов компании и предлагаются идеи относительно применения анализа данных для их совершенствования, формулируются конечные цели анализа. Для этого к обсуждению приглашается как можно больше заинтересованных специалистов и экспертов. Результатом этапа должен стать план аналитического проекта. Кроме этого, необходимо убедиться в целесообразности проекта, прежде чем тратить на него ресурсы.
- **Понимание данных** (Data Understanding). Данная фаза включает в себя более детальное изучение имеющихся данных. Ее цель — избежать непредвиденных проблем на стадии подготовки данных, которая, как правило, является самой сложной частью проекта. Начальное изучение данных предполагает организацию доступа к ним, их исследование с использованием таблиц и графиков, оценку качества данных и разработку соответствующей документации.
- **Подготовка данных** (Data Preparation). Является одним из наиболее важных и зачастую трудоемких этапов аналитического проекта, который может поглощать 50-

70% времени, усилий и ресурсов. В зависимости от специфики компании и направления ее деятельности подготовка данных обычно включает:

  - консолидацию данных;
  - формирование выборок;
  - агрегирование;
  - обогащение данных;
  - очистку данных;
  - разделение данных на обучающие и тестовые.

- **Моделирование** (Modeling). На данном этапе строятся и внедряются аналитические модели. Моделирование обычно проводится в несколько итераций. Сначала запускается несколько моделей с параметрами по умолчанию. Затем параметры настраиваются таким образом, чтобы модель выполняла требуемую обработку данных. Если это не удается, приходится возвращаться на этап подготовки данных и вносить изменения.
- **Оценка** (Evaluation). На этом этапе делается оценка того, соответствуют ли результаты проекта критериям успеха бизнеса. Этот шаг требует четкого понимания заявленных бизнес-целей, поэтому нужно обязательно привлекать к нему ключевых лиц компании, принимающих решения.
- **Внедрение** (Deployment). Внедрение — это процесс использования новых идей и знаний для повышения эффективности компании.



В настоящее время CRISP-DM де-факто является стандартом для разработки проектов интеллектуального анализа данных и обнаружения знаний.