THE CRISP-DM METHODOLOGY HANDBOOK

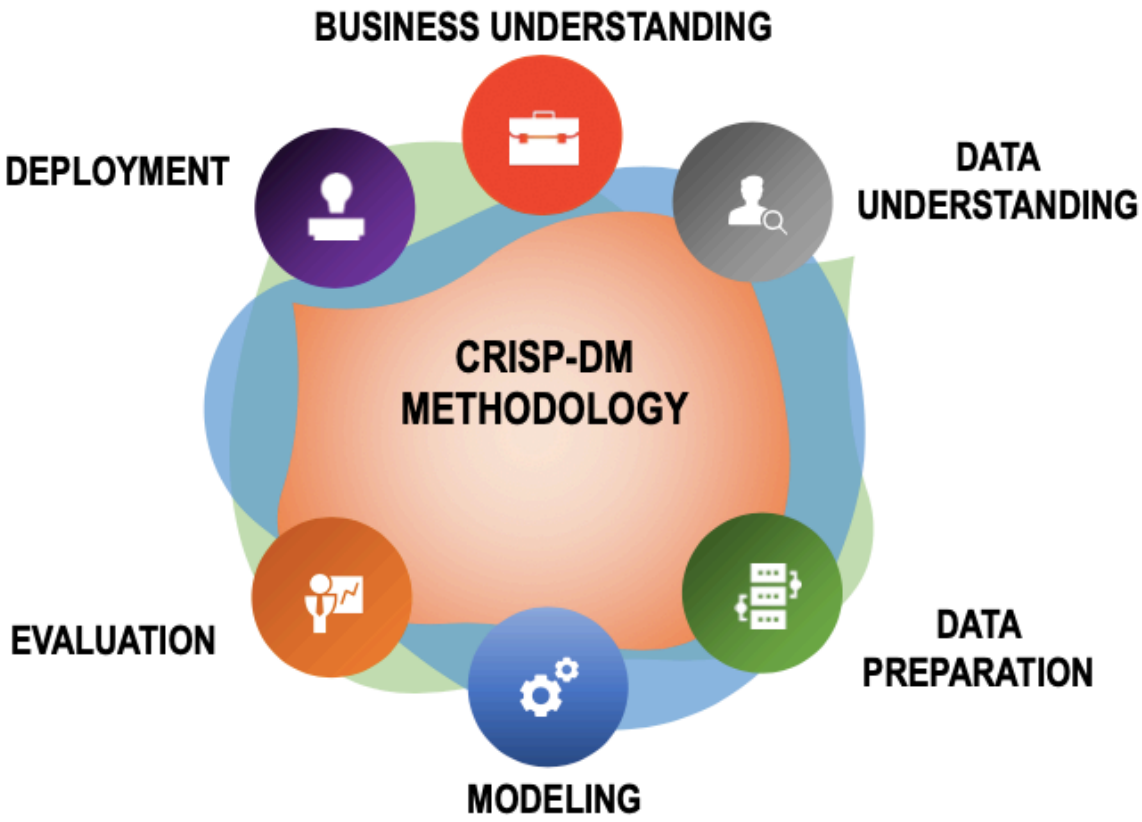
With 11 practical examples for deploying it in your e-commerce company

By Group F

Daniel Bilitewski
Guillermo Chacon Clericus
Mariana Narvaez Cubillos
Rabiga Shangereyeva
Jonas Hellevang
Nisrine Ferahi

Introduction

This guide can be used for you as a member of the analytical team of a company. We have presented a guide with examples for each phase of the CRISP-DM process, showing hands on which steps to take in each phase. The guide will teach you how to deal with each problem, and how to avoid mistakes or duplicate of work from any team that might be involved in the process. You will also be able to know how to deal with changes in the organization and its objectives, in data sources, and in available technology. This guide will walk you through the eight steps of CRISP-DM, their tasks and their outputs. These phases can be viewed as a cyclical methodology to emphasize the new challenges and way of work:



Source: Antonio Pita Lozano

1 – Business Understanding

Start the project by meeting with the team or department in your company for whom you are trying to solve a problem. If there is no team in your company that has a problem you are solving, maybe you are inventing a problem that is not actually relevant to the company. Make sure to always move from a business problem to a data problem. A good way to ensure this is by solving problems for other teams in your company, not a problem that your team came up with by itself. Furthermore, you need to make sure to define a business action which is taken based on the output of the model. This can happen both in an automatic system, for example an automatic email being sent to a customer, or in a manual way with a human in the loop who takes a decision based on a recommendation or insight from the model.

Next, make sure that the business objective is aligned with the overall strategy of the company. This is important to avoid having silos in the company, where different teams are working independently and are not aligned with the overall strategy.

Example 1: You are working with the purchasing team on optimizing the purchasing of a certain product category, to increase efficiency. However, the CEO has recently declared that product availability is the number one goal of your company for the next fiscal year. Of course, saving costs through more efficient purchasing is good, but maybe you can find a project more relevant to the overall company strategy.

After you have a well-defined business problem that you try to solve for a team in your company, continue by thinking about all other stakeholders involved in your solution. This includes both internal and external stakeholders. Internal stakeholders are other teams, which might be influenced by the model you are developing. Also, you should start considering at this stage if solving the business problem is possible with the existing infrastructure. External stakeholders are primarily customers, but also suppliers, delivery partners and other companies involved in the supply chain. Take strongly into consideration the ethical implications of your solution to the business problem, as well as legal constraints such as GDPR.

Example 2: You are working with team working on conversion optimization. Your goal is to create a model which will predict the likelihood along the purchasing process of a customer finishing the purchase. You should discuss with your IT department whether the required data from web analytics is currently being created and stored. If the right infrastructure doesn't exist, you might not be able to solve the business problem. You will then need to consider working on a different problem, or to ask managers to give a budget for supplying the required infrastructure first.

After you have a clear business goal and considered all implications on other stakeholders of solving the business problem you need to identify the most specific key performance indicator (KPI) that measures the success of the problem. Make sure to not only chose the right KPI, but to also have a clear understanding of how the impact of the model you are planning to build can be measure with this KPI. It is important to understand that for example increasing customer lifetime value (CLV) and increasing revenue is not the same thing. Even though a higher CLV leads to more revenue, it is a much more specific goal. Try to choose a specific goal as possible, while making sure that you chose a KPI that is easily measurable and that is relevant to other stakeholders in your company, so that you can communicate the impact of your model easily.

Example 3: Your goal is to develop a model which finds the right customer groups to target with a marketing campaign. You can have many kinds of metrics here: increasing revenue, increasing CLV, increasing click-through-rate of the email campaign, etc. Make sure your KPI is measurable and that you can determine the impact of the model. Different KPIs lead to

different outcomes. For example, increasing click-through-rate by 20% might not increase revenue because it leads to no sales. Was the model still successful in this case? Optimizing online advertisements might lead to only a 2% increase in CLV, but a 20% increase in revenue compared to the last campaign, because it was most relevant to new customers. Ultimately, choosing the metric has a big influence on evaluating the success of your model.

Finally, produce a project plan which outlines the goals of the project, how to measure its success in the context of the business through a KPI, what stakeholders are involved and what the timeline is. Most important here is to determine milestones in the project, especially if certain phases have dependencies on other work, maybe from IT, legal or another team.

Task and outputs of the phase:

- Determine Business Objectives:
    - Background Report analysis
        - Identify key actors in the project
        - Identify Sponsor of the project
    - Business Objectives
        - Describe the problem
        - Business Questions
        - Business Requirements
        - Benefits of the project
    - Success Criteria
        - KPI, measure criteria
        - Improvement Threshold
        - Identify who asses the increase
- Assess the situation
    - Inventory of Resources:
        - Base Hardware availability and usefulness
    - Requirement, Assumptions and Constrains:
        - All Requirements for the project
        - Legal, security, privacy constrains
    - Risk and Contingencies
        - Business, Organizational, Financial, Technical Risk
        - Data Risk
        - Contingency plan on risk
    - Terminology
        - Check Glossaries, or develop them
    - Cost and Benefits
        - Data Collection Cost
        - Solution implementation cost
        - Benefits
        - Operating cost
- Determinate data mining goals
    - Data Mining Goals
        - Business to Data Mining translation
        - Data problem Type
    - Success Criteria
        - Criteria for model
        - Benchmarks definition
- Produce project Plan
    - Project plan
        - Estimate the resources needed to achieve the plan
        - Identify critical steps
        - Make decision points
        - Review points

- Initial Assessment of tools and techniques
  - Selection criteria for tools
  - Potential tools and techniques

2 – Data Understanding

To understand which data could be relevant, first talk to the team for whom you are solving the business problem to ask them which kind of information is relevant from their perspective and where to find it. Next, internally in your team or together with other technical teams determine the kind what data is available in your company. If the business team suggested information that is not available, discuss with the technical team if this data can be gathered and what infrastructure would be required to do so. If the technical team is aware of certain tables in your database that could be relevant, share this information with the business team to get their opinion on the relevance of this data. This sharing of ideas helps to build a joint understanding of the problem you are trying to solve.

Example 4: You are working with the marketing team to improve the success of your campaigns. The technical team suggests using sentiment analysis (the feelings in a text) and semantic analysis (the meaning or topic of a text) to analyze social media data to better target potential customers. The marketing team wasn't previously aware of this technology and after sharing this information with them, they can start thinking about how to generally integrate this in the way they think about developing campaigns.

After getting a list of data that could be relevant to your problem, you need to assess which of this data can be used in your model. If the model is used in production outside the analytical tool, you need to make sure that the data can be available automatically, at the right time and in the right quality. This will maybe not be possible for all kinds of data or it might require new infrastructure.

Example 5: You are building a recommendation engine and together with the teams involved, came up with a long list of possible ways to measure the behavior of the customer on the website, in order to give better recommendation. But after talking to the website development team and the IT department, you realize that many of these aren't measured yet and that making them available in real-time is also not possible with the current IT infrastructure in your company. You will not be able to use both of this kinds of data in your model. However, make sure to develop a strategy and plan with both departments if this data should be collected in the future and how it can be collected.

Next, do an exploratory data analysis to understand the data better. In this process, you might find a lot of data errors, which you will clean to enable your analysis. Make sure to write a report on all data issues, and to directly discuss with the data owners on how to improve the quality of the data. The main goal of this step is to gain insights into how the data pertains to the problem you are trying to solve and how it can be used in the model. In this stage, you might also find insights that are relevant to the business and directly need to be communicated with the involved stakeholders.

Keep in mind that with each of these steps, the potential features you can use in your model becomes less and less, due to not being available, not being able to be used in production, or because you can't provide the data in the right quality. However, keep in touch with the stakeholders who can solve this problem so that it becomes part of the long term strategy of data acquisition and managing data as an asset. Maybe at a later point, the data is able to be made available and it will improve your model.

<u>Task and outputs of the phase:</u>

- Collect Initial Data
    - Initial data report
        - Plan needed information
        - Check availability of information

        - Select tables and attributes of interest
    - Data Description Report
        - Availability of data
        - Data types and values
        - Key relationships
        - Overlapping data
        - Assumptions
- Explore Data
    - Data Exploration report
        - Properties of data
        - Sub-populations
        - Perform hypothesis on data
        - Basic analysis to verify hypothesis
- Data Quality
    - Data Quality report
        - Check keys
        - Missing values/data
        - Format/typo/spelling mistakes
        - Outliers and deviations
        - Other types of mistakes

3 – Technology Platform

The solution you are planning will probably require some new infrastructure, which may be data storage, new ETL processes, data pipelines or analytical tools. The challenge is that there are many other projects in your company that also require new infrastructure. It's essential that there is a long-term strategy in your company which specifies the infrastructure that will be built and what technologies will be used.

Example 6: AWS might offer the perfect cloud solution to your problem. However, you company is using only Azure so far. Having a multi-cloud infrastructure might lead to integration problems down the line. Therefore, discuss with the IT department if it is allowed to use an AWS solution or if you have to find a solution offered by Azure.

If possible, for the data ingestion, storage, processing and serving, use technologies that are already available in your company. Make sure to use simple models and architectures to avoid having large dependencies on versions of certain libraries, programming languages or certain software. If there are too many dependencies in the data architecture of the company, it will become harder and harder to update the technologies.

If you need to license tools from vendors, make sure to account for the risk that the vendor might go out of business. You should develop a backup plan for how to maintain the infrastructure in such a case.

Output of this step:

- Document specifying the architecture and all tools required to deliver the solution
- Document specifying all dependencies and a rough plan on when to update to new versions
- If solutions from external vendors have been used, a plan on what to do if the vendor cannot support the solution anymore

4 – Data Preparation

First, you need to define the target that you are trying to predict. It is essential that the target is feasible to achieve the goal of the model. Furthermore, it is important that the performance of the model can be tracked over time. You need to choose the target in such a way that there is a feasible method to define a control group to track the accuracy of the model over time.

Example 7: You are building a model trying to predict which customer will be recurring customers and which are one-time customers. But what is a recurring customer? Someone that bought again in the next three months, or six months, or one year? This decision is crucial to define your target values. Also think about model performance tracking. If you define recurring customers as someone who has bought within the next 6 months, you need to wait six months to know the accuracy of your model. Is it really feasible to wait six months to test the accuracy of your model? Probably not.

Next, define how to clean data, such as dealing with missing values and outliers. There are two ways to do this: update the data permanently in the database or clean it in the data pipeline before inputting it into the model.

Define an accurate schema of how your data is organized and connected is essential to achieve a good understanding of it, you need to ensure that tables have the right primary and foreign keys and the structure of them is the correct one (primary keys must have unique values, foreign keys cannot have values that are not in the primary key).

Task and outputs of the phase:

- Select Data
  - Rationale for inclusion/exclusion
    - Select data that met conditions of selection
    - Sampling the data for model
    - Report of the decision
    - Make correlations on the available data
- Clean Data
  - Data Cleaning Report
    - Noise dealing
    - Special values address
- Construct Data
  - Derived Attributes
    - Normalization of data
    - Addition of new attributes
    - Dealing with missing values
- Integrate Data
  - Merged Data
    - Check availability of integration
    - Integrate the sources and tables
- Format Data
  - Reformatted Data
    - Rearrange data
    - Reorder data
    - Model specific changes

5 – Modeling

In the modeling phase, start off by specifying what kinds of models are suitable for your problem. You must consider the accuracy-interpretability tradeoff. If the output of the model needs to be understandable, maybe something like a logistic regression, KNN or decision tree is best. If only accuracy matters, then a Random Forest, Gradient Boosting Algorithm, Support Vector Machines or Neural Network might be best.

In the first iteration, make a very simple model to get a baseline of the performance you can expect. In certain situations, it is even recommendable to directly deploy this model if it is good enough, to get an impression of the challenges of deploying a model, before you invest a lot of time building something more complex.

Make sure to try multiple algorithms to compare their performance. You will need to find the right hyperparameters for each model to make the most out of it. A good method for this is grid search, were you specify a vector of different values to test for each hyperparameter. The model will then be trained with each combination of these hyperparameters, to find out which are best. If you do an initial grid search for each candidate model, you should have a good indication of which model will perform best for your problem. After choosing your model, you can do a more fine-tuned grid search to optimize it even more.

To find the best model to solve your business problem, accuracy might not be the best metric for a classification problem. Consider the impact of the false positive and false negative rate on your model performance and the action you take as a result of the model. This is especially important when training on data with an unbalanced target.

Example 8: You are making a model analyzing social media data to determine which fashion brands and products are popular at the moment, because you have products from many different brands on your website, but you want to know which is best to advertise. The higher the false positive rate, the more products and brands you advertise which aren't popular with your potential customers. This creates costs for marketing and lost revenues from losing potential customers. In this case, you need to consider making a model with a low false positive rate. It doesn't matter much if the false negative rate is, because this just means you won't advertise a product which is popular, but if you advertise another popular product instead, this is almost no negative impact. Therefore, for such a model, you want to optimize (i.e. reduce) the false positive rate, instead of the accuracy.

When training the model, make sure to split the data into a train, validation and test set. The train set is used to train the model and the validation set is used to evaluate the model on unseen data in order to avoid overfitting. Only after you have tested all hyperparameters in the grid search and decided on all final model you will evaluate it against the test set to know its performance on unseen data. Don't use the test set before, otherwise your model development will be biased to optimizing your model for the test set, even if you don't train the model on the test set, because you would only consider the score on the test set for choosing hyperparameters.

It is important to consider that not all features you have chosen in the data preparation step might be relevant and that you might need to create new features. Furthermore, certain features might work better with some models than with others. Make sure to experiment with this, but also consider that there is an infinite combination of features, models and hyperparameters which you can test, which makes it essential to plan on how to stay within the timeframe and budget of this project. Consider which is the best time to stop, because a 0.01 percentage point increase in accuracy might be hard to achieve at some point and will probably not add much value to your business.

Task and outputs of the phase:

- Select modeling technique
  - Modeling technique
    - Select the appropriate modeling technique
  - Modeling assumptions
    - Define the assumptions made by the technique
    - Compare the assumptions with the ones in the data report
    - Ensure that the assumptions align, otherwise reformulate the data preparation phase.
- Generate test design
  - Test design
    - Check the test design
    - Decide on sample size
    - Prepare test data
- Build Model
  - Parameters Setting
    - Set initial parameters
    - Justify the parameters selection
  - Models
    - Run the model on the database
    - Post model analysis of the data
  - Model description
    - Describe the characteristic of the model
    - Describe the interpretation of the results
- Asses the Model
  - Model Assessment
    - Evaluate results
    - Select best model
    - Interpret result in business terms
    - Check for reliability, plausibility repeatability
  - Revised Parameter Setting
    - Adjust parameters if necessary

6 – Evaluation & Results Presentation

Now that you decided on a final model, you need to meet with the team you are developing this model for to evaluate if the model solves the business problem as intended. It is crucial here to consider again the action which is taken as a result of this model. This can either be an automated action or it can be done by a human, as mentioned above. If the model can't properly trigger to action that solves the business problem, it is not the right model, even if it performs very well in terms of accuracy, F1-Score or any other metric you used to choose them model.

Next, consider the impact the model has on all stakeholders you defined in the business understanding step. In the process, while choosing the data and the model, things might have changed that created an unintended impact that you didn't consider before. It is important you make the relevant managers aware of any potential impacts and issues.

After this, review the entire model development process so far. This is essential to improve the data science operations in your company. When you find ways in which the process could have been improved, present them to your team and all relevant internal stakeholders. The team managers and product managers are then responsible to ensure that these improvements are implemented in all future projects across the organization. This is very similar to the continuous improvement processes already present in many companies. You can benchmark these continuous improvement processes to see what the best practices are for successfully implementing changes to your data science processes.

Example 9: After a few weeks into the development of your model, you realize that some of the data you are using for training won't be available when you want to productionize your model. This means that a lot of time and computational resources you invested into feature engineering and model development was wasted. How did this mistake happen? Was there a miscommunication with the IT department or another team? How can these issues be avoided in the future?

Finally, wrap up the documentation and make sure it is accessible to other people in your organization. Knowledge management is very important, so make sure that all knowledge is shared with the relevant decision makers, teams and data scientists so that someone else won't be trying to solve a problem in the future that you already solved today.

At the end, you will need to make a presentation to managers. Make sure to tell them what you did and not how you did it. They are usually not interested in the technical aspects of your solution and probably won't understand it. Make sure to adapt your presentation to your audience and to use language and metrics that they understand and that are important to them. Tell them not only the KPIs you are improving with your solution, but also the estimated business value in monetary units to clearly communicate your impact. When successfully finishing a project, it is also an opportunity to ask for more resources, infrastructure or something else that you need for future projects.

<u>Task and outputs of the phase:</u>

- Evaluate results:
    - Assessment of data mining results
        - Understand the results
        - Interpret the applicability of the results
        - Determine if new business objective can be obtained from this result
    - Approved Models:
        - Select the models that approved the criteria
- Review Process:

    - Review of Process
        - Overview of the data process
        - Analyze each step to see if they are adding value to the final result
        - Eliminate the steps that doesn't add value
- Determine Next Steps:
    - List of Possible actions
        - Analyze potential for deployment
        - Estimate potential for improvement
        - Check the resources to see if more iterations are possible
        - Recommend alternatives
    - Decision
        - Rank the actions
        - Select one of the options
        - Document selection reasons
- Produce Final report:
    - Final Report
        - Create final report
        - Identify target group for the report
        - Select findings to be included
    - Final Presentation
        - Decide target group for the presentation
        - Select items from final report to include in presentation

7 – Technical Deployment

Once the final model and results have been validated and it is time to insert them into the IT systems of the company, several decisions must be made during this phase.

The first thing to do is draw a conceptual model on how the current data and the new data will flow through the IT systems. This will provide a contextual view on how you should implement it and decide on the relevant factors that we will discuss next.

Given the model we need to decide how we will connect the model execution with the analytical tool. The two approaches are to do it in the same tool as the analytical or outside of it, each approach has its own advantages and disadvantages: Inside is easy to deploy and train but harder to integrate, and vice versa.

Example 10: You want to send a newsletter to your customers every Sunday with the offers you have in the upcoming week. A machine learning algorithm should decide which offers to send to which customers. Such a solution can be deployed within the analytical tool because there is no decision that must be made in real-time and the output of the model can be exported manually to the CRM system. If you wanted to deploy this model outside the tool, you would have significant costs for rewriting the tool in a different programming language, as the IT department doesn't use the language and libraries you used for modeling. In this example, this cost and effort can be avoided because it is not required.

Temporal planning should also be considered as a main factor in the technical deployment to consider what will happen with the model as time goes through and more data comes in, you should decide on how the model is updated, how much temporal data it uses and what it will do with the older models.

Finally, integration with other applications must be addressed in this phase, to understand how your model (inside or outside the analytical tool) connects to other resources (CRM, Web Pages, Apps, API, etc.) is key in the deployment phase.

<u>Task and outputs of the phase:</u>

- Technical Plan Deployment:
    - Deployment plan
        - Summarize results
        - Evaluate alternative deployment options
        - Decide on Monitoring factors
        - Decide how the model will be deployed within the IT systems
        - Identify potential problems and challenges

8 – Business Deployment

The final phase is all about taking everything that has been done and indebt it into the operation of the company, basically a model is a tool to make relevant decisions, now it's time to make those decisions.

As a first step is to decide who will be the users of this model, in this phase you must recall the sponsors, business areas that stated the problem and key actors that you analyzed in the Business understanding phase of this methodology. Once you have successfully identified the key final users of this models you should produce all the documentation and platforms needed for the model to be used by them, this includes but not limited to: Documents and Reports, Infographics, Dashboards, Score Cards, etc.

A key factor in this phase is to decide actions aligned with the threshold of the model, this decision should be correlated with the business problem described in the previous section.

Last thing is to plan on the Monitoring and Maintenance of the plan, as business and time is changing, so is the plan, so several decisions on when to monitor, explore, re-estimate and re-train the model are necessary. Several metrics can be developed to address this, check the variable stability, the model stability over time and the capability of it.

Example 11: You are again on the project where you want to send newsletters with relevant offers to customers. How does this integrate with other business units? Maybe you can make slightly worse offers but use this as an opportunity to reduce excess stock in your warehouse. How do you monitor the relevance to customers? You can consider click-through rates, purchases rates and other metrics. What is the timeframe of the model usage? You need to decide on how many weeks you use it before remodeling or retraining it. Consider also that at some point the business strategy might change. If consumers receive emails too often, they will get annoyed. Furthermore, customers only have limited spending power. Your model can predict very relevant recommendations to customers, but if the business execution is wrong, it will still perform badly. Therefore, consider also how to evaluate your business decisions and its impact on the model.

Task and outputs of the phase:

- Business Plan Deployment:
    - Deployment plan
        - Summarize results
        - Evaluate alternative plans
        - Decide on the knowledge acquired by the model
        - Decide on how this knowledge will affect the decision process for the user
        - Decide on how the model will be used and monitored
        - Identify possible problems