**THE DATA MINING PROCESS…**

Data Mining is not something that is done once and then forgotten…it is an ongoing cycle of execution and discovery…any given project may address just a small portion of a much larger problem or project…and it is possible that data scientists will be called on to work on additional components of that project or to go back to the original project as business needs change…

Data Scientist (like programmers and network engineers and every other segment of technology) revisit their past projects to see whether the models developed are still effective, and to look for opportunities to improve on what has already been done, or to apply the same principles and techniques to current or future projects…

Like most everything else in Information Technology, the development process has no ‘official’ rules…there are ‘guidelines’ that are adjusted based on each organization’s own goals and experiences…and as you will see (or have already seen), every organization has its own standards and guidelines that they follow…flexibility is key to allow for changes based on any given scenario’s own unique goals or requirements…

But at their heart, most every guideline has the same core items that are needed to identify, justify, and execute a project, whether the project be an application development project or a data science project…

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is the dominant process framework for data science projects…it is an ‘open standard’…that means it is a ‘guideline that anyone can use or adapt to meet goals and requirements…we will be utilizing this framework in the development of your data science project…

**What Exactly is CRISP-DM?...**

The CRISP-DM process model is a step-by-step approach to data mining…it was created BY data miners FOR data miners…it was conceived in 1996 for a project that was let by five companies: Integral Solutions Ltd (ISL), Teradata, Daimler AG, NCR Corporation and OHRA, an insurance company…each company brought their own unique experiences to the project…the first version of the methodology was published in 1999…over the next ten years, discussions and updates were made…at that point, the founding organizations all moved on to other projects and no ‘official’ changes have been made to the process since 2009…today, IBM is the primary corporation that currently uses the CRISP-DM process model…

CRISP-DM defines the data-mining process primarily from a **business** perspective…it defines what needs to be done without the technical details of how…

CRISP-DM is still the most widely used form of data-mining mode…it is industry, tool, and application neutral…it contains many of the processes found in application development but takes a ‘data’ view of them…with the exception of a basic project plan, the model does not specify project management activities the way that a formal ‘scoping document’ does (remember those?)…those tasks still need to be performed but are specified and maintained by the project manager…

CRISP-DM attempts to provide structure to the data mining process while at the same time leaving users free to make their own decisions about data approaches and specific technical decisions…

**CRISP-DM Process Phases…**

The CRISP-DM Process Model has six primary phases…

1. **Business Understanding:** focuses on understanding objectives and requirements of the project…used to get a clear understanding of the problem to solve, how it impacts the organization, and the goals for addressing it…contains tasks that are foundational project management activities…
2. **Data Understanding:** focuses on identifying, collecting, and analyzing data sets that help accomplish project goals…reviews the data, documents it, and identifies data management and data quality issues…
3. **Data Preparation:** prepares the final data set(s) for use in modeling…
4. **Modeling:** builds and assess various technical models to identify patterns within the data…
5. **Evaluation:** reviews the models used and the patterns discovered in the ‘Modeling’ phase, and assesses their potential for business use…
6. **Deployment:** identifies how models can be implemented and accessed for everyday business use…

A screenshot of a computer

AI-generated content may be incorrect.

Each phase contains several major tasks…each task has at least one deliverable…the deliverables are primarily reports summarizing work completed, and documenting information learned…

**IMPORTANT NOTE :** while many organizations may have standard templates for each deliverable, the CRISP-DM process does **NOT** provide templates for these deliverables…you will have to do a little research or use prior projects as a guide for constructing your deliverables…it is key that your documents be CONSISTENT in terms of structure and presentation (note how all of my Power Points have the same ‘look and feel’)…

**Summary of Data Mining Project Tasks and Deliverables…**

Below is a summary of the phases, tasks and deliverables that are needed to successfully execute a Data Science Project…each of these items will be reviewed and discussed in more detail in the following pages…

Yes!...it LOOKS like a lot of items…but as you will see, each item must be done regardless of whether or not you take the time to document everything…

* **PHASE 1.0 : Business Understanding…**
  + Task 1.1 : Determine Business Objectives…
    - Deliverable : Project Scope Document – Part 1…
  + Task 1.2 : Assess the Situation…
    - Deliverable : Project Scope Document – Part 2…
  + Task 1.3 : Determine Data-Mining Goals…
    - Deliverable : Data-Mining Scope Document…
  + Task 1.4 : Produce a Project Plan…
    - Deliverable : Data Mining Project/Resource Plan…
* **PHASE 2.0 : Data Understanding…**
  + Task 2.1 : Gathering Data…
    - Deliverable : Data Collection Report…
  + Task 2.2 : Describing Data…
    - Deliverable : Data Description Report…
  + Task 2.3 : Exploring Data…
    - Deliverable : Data Exploration Report…
  + Task 2.4 : Verifying Data Quality…
    - Deliverable : Data Quality Report…
* **PHASE 3.0 : Data Preparation…**
  + Task 3.1 : Selecting Data…
    - Deliverable : Data Rationale Report…
  + Task 3.2 : Cleaning Data…
    - Deliverable : Data Cleansing Report…
  + Task 3.3 : Constructing Data…
    - Deliverable 1 : Data Attribute Report…
    - Deliverable 2 : Data Generation Report…
  + Task 3.4 : Integrating Data…
    - Deliverable : Merged Data Set…
  + Task 3.5 : Formatting Data…
    - Deliverable : Final Formatted Dataset…
* **PHASE 4.0 : Modeling…**
  + Task 4.1 : Selecting Modeling Techniques…
    - Deliverable 1 : Defined Modeling Technique(s)…
    - Deliverable 2 : Defined Modeling Assumptions…
  + Task 4.2 : Designing Tests…
    - Deliverable : Test Design Document…
  + Task 4.3 : Building Model(s)…
    - Deliverable 1 : Parameter Definitions…
    - Deliverable 2 : Model Descriptions…
    - Deliverable 3 : Data Models…
  + Task 4.4 : Assessing Model(s)…
    - Deliverable 1 : Model Assessment…
    - Deliverable 2 : Revised Parameter Settings…
* **PHASE 5.0 : Evaluation…**
  + Task 5.1 : Evaluating Results…
    - Deliverable 1 : Result Assessment…
    - Deliverable 2 : Model Approval…
  + Task 5.2 : Reviewing the Process…
    - Deliverable : Process Evaluation Report…
  + Task 5.3 : Determining the Next Steps…
    - Deliverable 1 : Possible Actions…
    - Deliverable 2 : Final Decision…
* **PHASE 6.0 : Deployment…**
  + Task 6.1 : Planning Deployment…
    - Deliverable : Deployment Plan…
  + Task 6.2 : Planning Monitoring and Maintenance…
    - Deliverable : Monitoring and Maintenance Plan…
  + Task 6.3 : Reporting Final Results…
    - Deliverable 1 : Final Report…
    - Deliverable 2 : Final Presentation…
  + Task 6.4 : Review Project…
    - Deliverable : Team Experience Assessment…

**A Note on Documenting Project Work…**

You might have noticed from the above that much of the CRISP-DM process model is focused on reports and documentation…the documents are the means of preserving information about what was done and how it was done that you and others will not have to decipher things later on…

What may seem obvious to the people working a project day in and day out, is not so obvious to new people coming onto a project or management that wants to review project details…consider if you moved on to a new project and were given no documentation…you would have to spend a LOT of time trying to figure out what items are, and how they were created, and why certain decisions were made…you might repeat work or make your team repeat work, and you might make the same mistakes that the originators of the project made…and, failure to document reasons for making certain decisions, or to proof that data privacy obligations were met, could have legal consequences…

Yes, documenting is a pain…yes, you would rather be doing the ‘fun’ stuff…but you will find that documenting is key to the success of a project and to your future as a data scientist…

**Details of Data Mining Project Tasks and Deliverables…**

The rest of this document contains the details of what every task and deliverable is…it is expected that you perform every task and submit every deliverable in a consistent and timely manner…it is also to be expected that as the project progresses, you may have to go back and adjust prior deliverables…for example, if you discover in Phase 3.0 that the data is inadequate, you may have to go back to Phase 2.0 to reexam sources…the earlier you discover issues and problems, the easier it is to resolve them…

Again…yes!...it LOOKS like a lot of items…and yes, some deliverables have a lot of information (e.g., Task 3.3 Data Attribute Report)…but as you will see, each item must be done regardless of whether or not you take the time to document it…and none of these documents require pages and pages (and pages) of writing)…be thorough…but get to the point with as little ‘prose’ as possible…

**PHASE 1.0 : Business Understanding**

As with everything else we do, we need to define goals and objectives BEFORE we start so that we know where we are going (“If one does not know to which port one is sailing, no wind is favorable” - Lucius Annaeus Seneca)…

“Business Understanding” includes four tasks…as you will see, tasks 1, 2, and 4 are foundational project management activities…the four tasks are:

1. Determine Business Objectives…
2. Assess the Situation…
3. Determine Data-Mining Goals…
4. Produce a Project Plan…

**Task 1.1 : Determine Business Objectives…**

**Deliverable : Project Scope Document – Part 1…**

Before you can begin to determine the data needed or the tools or visualizations needed, you need to understand the goals, objectives, and requirements of a project…you also need to identify the business reasons for wanting to achieve the goals stated, and what the success criteria would be…a clear understanding is needed of:

* The problem management wants to address…
* The business goals…
* Business success criteria…
* Constraints (e.g., time, money, resources, technology)…
* Organizational and business impact (e.g., how the problem and possible solutions fit in with the business)…

***Deliverables*** for this task include five items (usually brief reports focusing on just the main points):

* **Background / the Problem Management Wants to Address :** Explain the business situation that drives the project…this is typically a few paragraphs to a few pages…for example:

*Our client, a regional planning commission, seeks to influence property use to enhance the quality of life for local residents. The planning commission has a broad charter which allows it to consider wide-ranging issues including employment, recreation, environment, and many other aspects of community life; however, the commission’s role is purely advisory. It has a great deal of latitude to select issues for study, conduct research, and make policy recommendations to local lawmakers and staff, but does not have independent power to set regulations or influence property owners. Commission members (and others in local government and civic organizations) believe the best opportunity to influence property use occurs when the property changes hands. This implies that local government planning efforts can achieve greatest impact by focusing on properties which are about to change ownership. This poses a problem: the best time to act is before property changes hands, but the local government does not have dependable information about which properties are likely to be transferred. (Commercial real estate listings may be useful, but they do not cover all property transfers, and the best time to act may be before the property is listed.)*

*Earlier research has identified a number of factors believed to indicate impending change of ownership; these include nonlocal ownership, multiple building code violations, and foreclosure, among others. While commissioners have good reason to believe that these factors influence the likelihood of property to change hands, their effects have not been quantified.*

* **Business Goals:** Define what the organization intends to accomplish with the project…this is usually a broader goal than the actual data mining project…for example, the business goal might be to increase sales from a holiday ad campaign by 10 percent year over year…
* **Business Success Criteria:** Define how the results will be measured…these need to be clearly defined ***quantitative*** success criteria…qualitative or subjective criteria can also be used (e.g., “gain insight”, “get a handle on”)…but it must be defined up front who will determine if the criteria have been fulfilled…
* **Constraints : *BEFORE*** any work is started, any known limitations need to be defined…this includes limitations on staffing, technology that can be used, timeframes for completion, or money that can be used for new purchases of hardware or software…any of these limitations could have an impact on potential mining strategies and solutions…
* **Organizational and Business impact :** all projects are undertaken because of the impact they will have on the organization (both positively and negatively)…projects are often approved (or not approved) based on the potential impact they might have on an organization and how it operates…

**Task 1.2 : Assess the Situation…**

**Deliverable : Project Scope Document – Part 2…**

Now that the objectives are defined, there needs to be a more in-depth analysis on the issues and costs associated with the business defined business goals…this includes deeper explanations of project requirements, risk assessments, contingencies, and the availability of resources to actually complete the project…a cost-benefit analysis must also be conducted…

***Deliverables*** for this task include five in-depth reports:

* **Inventory of Resources :** A COMPLETE list of resources needed and available for the project…resources include people (e.g., data scientists, business users with expert knowledge of the business problem, data managers, technical support); as well as, data, hardware, and software…
* **Requirements, Assumptions, Constraints :** Although constraints were listed in Task 1.1, those were general business related constraints…these requirements, assumptions, and constraints are specific to the actual project…this will any defined completion dates, legal and security obligations for the data or its usage, and definitions for acceptable finished work…it is also here that you determine if you can have access to the data that you believe you will need…
* **Risks, Contingencies :** Identify causes that could delay completion of the project, and prepare a contingency plan for each of them…for example, what happens if there is an internet outage in the office?...what happens if data becomes unavailable?...
* **Terminology :** Never assume that everyone working on a project has the same depth of understanding of terms used in the business or in data mining…as part of project documentation, create a list of business and data-mining terms and definitions that are relevant to the project…including examples is always a very good idea...
* **Costs and Benefits :** EVERY project needs a cost-benefit analysis…this will tie together the cost of the project with the perceived organizational impacts…keep in mind that data mining projects are a little different than typical application development…the goal of most (not all) application development projects is to implement products and services to create revenue…the goal of some (not all) data mining projects is to obtain information from which other business decisions can be made…but even so, a cost-benefit analysis is ALWAYS required…

**Task 1.3 : Determine Data-Mining Goals…**

**Deliverable : Data-Mining Scope Document…**

In addition to defining the ***business*** objectives, ***data mining*** objectives also need to be defined…and BOTH sets of objectives must be in sync with each other…for example, if the business goal is to reduce customer attrition, the data-mining goal might be to identify attrition rates for various customer segments, and develop models to predict customers at greatest risk…

***Deliverables*** for this task include two reports:

* **Data-Mining Goals :** Define data-mining deliverables, such as models, reports, presentations, and processed datasets…
* **Data-Mining Success Criteria :** Define the data-mining technical criteria necessary to support the business success criteria…define these in quantitative terms (e.g., model accuracy, predictive improvement compared to an existing method)…qualitative criteria need to identify the person responsible for making the assessment…

**Task 1.4 : Produce a Project Plan…**

**Deliverable : Data Mining Project/Resource Plan…**

EVERY project needs a project plan…the project plan is the management tool used to identify tasks, and determine if a project is on time and on budget…

***Deliverables*** for this task include two reports:

* **Project Plan :** Outline the step-by-step action plan for the project…as with all project plans, include start and end dates, required resources, inputs (e.g., data, meeting with subject matter experts), outputs (e.g., cleaned data, model, report), and dependencies…explicitly state that certain steps may need to be repeated (e.g., modeling and evaluation)…
* **Initial Assessment of Tools and Techniques :** Now that you understand the business needs and requirements, you should have a good idea of the required capabilities needed to meet the data-mining goals…you should also have a good idea of the tools and resources that will be needed to complete the project…as with all projects, any new information could cause a change in the tools and resources needed…and if/when that happens, all changes will need to be approved…

**PHASE 2.0 : Data Understanding**

The second phase adds to the foundation of the Business Understanding phase…this phase focuses on identifying, collecting, and analyzing data that can be used to accomplish the project goals…you might also identify issues that can cause an update to the business understanding and goals, and in turn, the project plan..

“Data Understanding” includes four tasks…the four tasks are:

1. Gathering data…
2. Describing data…
3. Exploring data…
4. Verifying data quality…

**Task 2.1 : Gathering Data…**

**Deliverable : Data Collection Report…**

Now that a project has been defined and approved, there is one very important thing that we need to be successful…that is, THE DATA…and just any data…but the RIGHT data…without the right data, everything else falls apart…

There is ‘just’ one ***deliverable*** for this task - the initial data collection report…that means, we need to acquire the necessary data and (if necessary) load it into a data analysis tool…

Even though there is only one deliverable in this task, there are many items that need to be done…you need to verify that you have acquired or gained access to the data, tested the data access process, and verified that the data exists…you also need to load at least a sample of the data into any tools being used for data-mining to verify that the tools are compatible with the data…

All of the above needs to be completed before you can write this report…in order to complete these items, you need to plan, as follows:

* **Outline Data Requirements :** Create a list of the types of data necessary to address the data mining goals…include details such as time range and data formats…
* **Verify Data Availability :** Confirm that the required data exists, and that it is available to be used…if some data is unavailable, determine how to address the issue…if you consider other alternatives (e.g., substituting with an alternative data source, narrow the scope of the project, gathering new data), be sure that the business and all team members are in sync with that decision…
* **Define Selection Criteria :** Identify all data sources (e.g., databases, files, documents), and the specific tables, fields, and case ranges relevant to the project…

Once the above have been identified, you must actually obtain the data…import the data into the data-mining platform being used for the project to confirm that it is possible to do so and that you understand the process…you may also discover issues such as software or hardware limitations you had not anticipated )e.g., limits on the number of cases or fields, or on the amount of memory used), an inability to read source data formats, or difficulty dealing with data imperfections (e.g., incomplete dataset)…

Finally, after all of the above has been completed, summarize the gathering process in a report…the report should describe your requirements, and explain in some detail exactly what data was gathered and from what sources…also, confirm the data was actually obtained and is compatible with the data-mining platform being used…explain any difficulties or issues you had and how they were addressed…

Again…the deliverable for this task is a ‘simple’ report that requires a LOT of preparation…data access can be one of the most challenging (and frustrating) parts of the data-mining process…

**Task 2.2 : Describing Data…**

**Deliverable : Data Description Report…**

Now that we have gathered the data, we need to ***create a report*** that describes the data…this report will include the source and formats of the data, the number of cases, the number and descriptions of the fields, and any other general information that may be important…it also contains a brief evaluation of the suitability of the data for your data-mining goals…for example, verify that the data includes the expected fields and that there are sufficient cases for analysis…

**Task 2.3 : Exploring Data…**

**Deliverable : Data Exploration Report…**

Now that we have loaded the data and verified that it is what we need, we need to exam the data more closely…for example, look at the range of values and distribution of each variable…

Data exploration supports several purposes:

* Get familiar with the data…
* Spot signs of data quality problems…
* Set the stage for data preparation steps…

The ***deliverable*** for this task is the data exploration report. It’s the place to document any hypotheses or initial findings that you have developed during data exploration. This report should include a more detailed description of the data than the data description report, including distributions, summaries, and any signs of data quality problems.

**Task 2.4 : Verifying Data Quality…**

**Deliverable : Data Quality Report…**

Lastly…now that we examined the loaded data, we need to determine if it is good enough to support the goals of the project…minor data quality issues can often be addressed without impacting the project…but if the data quality is so poor that it cannot support the project, you will need to look for alternate data, and in some cases, alter or cancel the project…

Some examples of major data issues include:

* The data needed does not exist…
* The data exists but it is not available (e.g., it is restricted, there are legal issues with the data)…
* The data has severe quality issues (e.g., lots of missing or incorrect values that cannot be corrected)…

The ***deliverable*** for this task is a Data Quality Report…the report summarizes the data, identifies both minor and major quality issues discovered, and possible remedies or alternatives for quality issues…if there are serious data quality issues with no adequate solution, the report should recommend changes to the project goals or plans…any such changes would need to be approved by the project sponsor before moving forward…

**PHASE 3.0 : Data Preparation**

Most data-mining activity takes place in Phase 3.0, Data Preparation…here, all the data that was collected and reviewed in Phase 2.0 is further refined for use in modeling and visualization…

The “Data Preparation” phase includes five tasks…the five tasks are:

1. Selecting data…
2. Cleaning data…
3. Constructing data…
4. Integrating data…
5. Formatting data…

**NOTE 1 :** CRISP-DM does not *explicitly* mention datasets as deliverables for each of the data preparation tasks…but datasets need to exist and be properly archived and documented in order to move forward with each task…so although there is not a “one-to-one” correspondence between tasks and datasets, dataset information needs to be included in each deliverable report…

**NOTE 2 :** Although the five tasks are listed in *logical* sequence, the activities might actually occur in a different *physical* sequence…they may also be done many times…

**Task 3.1 : Selecting Data…**

**Deliverable : Data Rationale Report…**

In prior phases, we collected and verified data…now we need to determine SPECIFICALLY which portions of the data collected will be used for data mining…

The ***deliverable*** for this task is the rationale for inclusion and exclusion…the report will explain what data will, and will not, be used for further data-mining work, and the reasons for including or excluding each piece of data…the reasons should be based on relevance to project and data mining goals, data quality, and technical issues which were identified in Task 2.1…

**Task 3.2 : Cleaning Data…**

**Deliverable : Data Cleansing Report…**

Most data used in data mining (and everywhere else) is unlikely to be perfectly clean…in reviewing the data you may need to track down sources to make specific data corrections, exclude some cases or individual data or records, or replace some data with default values…you might also choose to use only a subset of the data for all or some of the data-mining work…

The ***deliverable*** for this task is the data-cleaning report…this report documents (IN EXCRUCIATING DETAIL) every decision and action used to clean your data…the report should cover and refer to each data quality problem that was identified in Phase 2.0…finally, it should address the potential impacts of your choices on the project or the final results…

**Task 3.3 : Constructing Data…**

**Deliverable 1 : Data Attribute Report…**

**Deliverable 2 : Data Generation Report…**

One result of the data cleansing process could be the need to derive new fields…for example, you could use an order and a delivery to determine how long a customer had to wait to receive their order…you may also need to aggregate data (e.g., total customers in New York), or create a new form of data (e.g, convert dates to Julian)...

***Deliverables*** for this task include two reports:

* **Derived Attributes :** This report describes what new fields were constructed, how they were constructed, and why…
* **Generated Records :** This report describes what new rows were constructed, how they were constructed, and why…

REMEMBER THE EARLIER NOTE : Although these tasks are listed in *logical* sequence, the activities might actually occur in a different *physical* sequence…they may also be done many times…

**Task 3.4 : Integrating Data…**

**Deliverable : Merged Data Set…**

The data used for data mining could exist in many different datasets…some or all of the data needs to be merged so that it can be modeled…

The ***deliverable*** for this task is the merged data set…as part of this, you also need to document how the merge was performed and why…

**Task 3.5 : Formatting Data…**

**Deliverable : Final Formatted Dataset…**

Data may come in various formats that need to be made consistent, or that may or may not be best for modeling…format changes are usually driven by the tools you are working with…once the data is merged, format and convert the data…waiting for the end helps to ensure that you are formatting all the data consistently…

The ***deliverable*** for this task is the reformatted data and (as you probably already guessed), a report that describes the changes made and why…

The Data Preparation Phase should end with a dataset ready for modeling and a *thorough* report describing the dataset…

**PHASE 4.0 : Modeling**

Finally, finally, finally!...

While this is the part that data miners like the best, you should be able to see how there would be no way to do a proper data model without completing the tasks outlined in Phases 1, 2, and 3…yes, you could have started here (this is technology – you can do anything!)…but the same way that coding without a flowchart can lead to omissions and many ‘do-overs’, all the prep work we did will help to insure that we start the modeling phase exactly where we need to be with minimal (if any) omissions and ‘do-overs’…

So now that the data has been prepared, we can start to look for patterns in the data that help us meet the stated goals of the business and the project…

The Modeling Phase includes four tasks…these tasks are :

1. Selecting modeling techniques…
2. Designing test(s)…
3. Building model(s)…
4. Assessing model(s)…

**Task 4.1 : Selecting Modeling Techniques…**

**Deliverable 1 : Defined Modeling Technique(s)…**

**Deliverable 2 : Defined Modeling Assumptions…**

As many of you already know, there are MANY different modeling techniques and visualizations…but not all of these will meet the needs of the project or highlight the results of your data mining in a way that makes sense for the business…

Now is the time to review the various techniques and determine which ones are best for the project…this decision is based on the kinds of variables being use, techniques available in the tools, and any business considerations important to the project…for example, some organizations favor methods with output that is easier to interpret…based on that, using a decision trees would be acceptable, while a neural network would not be…

***Deliverables*** for this task include two reports:

* **Modeling Technique(s) :** Based on the detailed knowledge we know have of the business goals and of the data being used, we should be able to determine the technique(s) to use for modeling our data…
* **Modeling Assumptions :** Some modeling techniques are based on certain assumptions…for example, some models are intended for use with data that has a specific type of distribution…you guessed it…you need to document any of these assumptions in this report…

Remember…NO ASSUMPTION IS TOO BASIC OR TOO SIMPLE TO NOT PUT DOWN…better too much than too little…

**Task 4.2 : Designing Tests…**

**Deliverable : Test Design Document…**

Once the model techniques have been determined, we need to identify how you are going to test the model to ensure that your model works…this includes not only identifying what you are testing and how, but also the approach to your data…for example, you might need to split the data into training, test, and validation sets…what percentages will you use?...will you vary the percentages?...

The ***deliverable*** for this task is your test design…the document should include your training, testing, and validation strategy, and how you plan to avoid introducing bias into the data…

**Task 4.3 : Building Model(s)…**

**Deliverable 1 : Parameter Definitions…**

**Deliverable 2 : Model Descriptions…**

**Deliverable 3 : Data Models…**

Most people believe that building models is the major part of a data miner’s job…but in fact, it is one of many (as you have seen)…in fact, the building of a model could consist of only a few lines of code!...but even so, addressing business goals through modeling IS the heart of the data-mining profession…

***Deliverables*** for this task include three items:

* **Parameter Settings :** Settings vary based on the model you are building and the tools you are using to build them…most tools give you the option of adjusting settings…each setting and each adjustment has an impact on the structure of the final model…document ALL settings (even if you did not change them), and document the reasons for changing or not changing a setting…
* **Model Descriptions :** Document the type of model used and why you choose it…include all variables used and explain why you choose them…explain how the model should be interpreted…document any difficulties encountered in the modeling process, or any important items of note…
* **Models :** At long last, the models themselves!...

**Task 4.4 : Assessing Model(s)…**

**Deliverable 1 : Model Assessment…**

**Deliverable 2 : Revised Parameter Settings…**

In Task 4.3, models were created and run…in Task 4.4 the models are reviewed from both a technical standpoint and a business standpoint, as to their applicability the project goals…this involves input from all team members, especially data scientists and the business users…

***Deliverables*** for this task include two reports:

* **Model Assessment :** This provides a summary of the information developed in the models…if several models were created, results should be ranked based on the assessment of their value to the goals of the project…the focus here is a TECHNICAL assessment of the model, NOT a BUSINESS assessment…the business assessment is conducted in Phase 5.0…
* **Revised Parameter Settings :** Based on the results of the model assessment, you may decide to fine-tune some of the parameter settings used to build the model…if parameters are changed, another round of modeling is performed, results evaluated, and the cycle begins again…

Often, data scientists just starting out leave the default parameter settings…in fact, you might not even notice options unless you look for them…as you gain experience, you will continue to experiment with parameters and how to use them…options vary based on the type of model and the tools being used…

**PHASE 5.0 : Evaluation**

Whereas Task 4.4 - Assess Model(s) in the Modeling phase focuses on technical model assessment, the Evaluation Phase looks more broadly at which model best meets the business needs…based on this evaluation, a determination must be made as to next steps…

In addition to evaluating models, the process used to create the models should also be evaluated…

The Evaluation Phase includes three tasks…these tasks are:

1. Evaluating results…
2. Reviewing the process…
3. Determining the next steps…

**Task 5.1 : Evaluating Results…**

**Deliverable 1 : Result Assessment…**

**Deliverable 2 : Model Approval…**

Now it is time to assess the value of the models for meeting the business goals that started the entire process…you need to identify reasons why the model would or would not be satisfactory for business use…the model should be tested in a practical application, to determine whether it works as well in the workplace as it did in your tests…

***Deliverables*** for this task include two items:

* **Result Assessment :** This assessment deals with how well the model results match the business goals…results should be summarized with respect to the business success criteria that was established in Phase 1.0 – Business Understanding…it is important to EXPLICITLY state if the business goals defined at the start of the project have been met...
* **Model Approval :** This is the formal statement of which (if any) models meet the business success criteria…

**Task 5.2 : Reviewing the Process…**

**Deliverable : Process Evaluation Report…**

Every project that is worked on should take the time to evaluate the process and the steps taken…the first four phases dealt with exploring data and developing models…now that the models have been assessed, it is time to look back and assess the PROCESS…was anything overlooked?...were all steps properly executed?...was enough time given to each step?...if issues or problems are identified that could impact the final product, they should be correct before deployment…

Summarize findings and correct anything if needed…it is an opportunity to identify areas issues that you might have overlooked and that might draw your attention to flaws in the work that you’ve done while you still have time to correct the problem before deployment…also consider ways that you might improve your process for future projects.

The ***deliverable*** for this task is a Process Review Report…the report should outline the review process, summarize the findings, and highlight any concerns that might require immediate attention (e.g., steps that were overlooked or that should be revisited)…changes to the process for future projects should also be identified…

**Task 5.3 : Determining the Next Steps…**

**Deliverable 1 : Possible Actions…**

**Deliverable 2 : Final Decision…**

The evaluation phase concludes with recommendations for what happens next…is the model ready to deploy?...should certain steps be repeated to improve the results?...should new data-mining projects be undertaken?...

***Deliverables*** for this task include two items:

* **List of Possible Actions :** Describe each possible next step, and provide reasons for and against it…
* **Final Decision :** State the final decision on each possible action, and provide the reasoning behind the decision…

**PHASE 6.0 : Deployment**

So why did we do all of this?...so that a customer can access the results to positively impact the business and improve the way the business is run…nothing else matters!...

The Deployment Phase has four tasks…these tasks are:

1. Planning deployment (your methods for integrating data-mining discoveries into use)…
2. Planning monitoring and maintenance…
3. Reporting final results…
4. Reviewing final results…

**Task 6.1 : Planning Deployment…**

**Deliverable : Deployment Plan…**

When the model is ready to use, a strategy is needed for putting it to work in your business…

The ***deliverable*** for this task is the actual deployment plan…this is a summary of the strategy for deployment, the steps required, and the instructions for carrying out those steps…

**Task 6.2 : Planning Monitoring and Maintenance…**

**Deliverable : Monitoring and Maintenance Plan…**

Data-mining work, like application development, is a cycle…deployment and monitoring usage are ongoing items that need involvement from all people involved…

The ***deliverable*** for this task is the monitoring and maintenance plan…this is a summary of the strategy for ongoing review of the model’s performance, and for updating models based on model performance or changes in available data…

**Task 6.3 : Reporting Final Results…**

**Deliverable 1 : Final Report…**

**Deliverable 2 : Final Presentation…**

Time to wrap things up!...

***Deliverables*** for this task include two items:

* **Final Report :** The final report summarizes the entire project by assembling all the reports created up to this point, updating those reports where needed, and adding an overview summarizing the entire project and its results…
* **Final Presentation :** A summary of the final report is presented in a meeting with management…it is also an opportunity to address any open questions or concerns…

**Task 6.4 : Review Project…**

**Deliverable : Team Experience Assessment…**

Now that the model has been deployed, it is time for the data-mining team to take a step back, and have a candid discussion as to what worked well, what did not work well, what could have been better, what should be avoided, and where can improvements be made in the future...

This deliverable is for use by the data mining team…it is a Team Experience Assessment…this document should outline any work methods that worked particularly well, so that they are documented to use again in the future…it should also identify any improvements that might be made to the process.it should document problems and bad experiences, with recommendations for avoiding similar issues in the future (you should NEVER identify a problem without also identifying potential solutions or mitigations in the future!)…

**NOTE :** While this deliverable is for the Data Mining Team that worked on a particular project, it is always a good idea to share your experiences with other Data Mining Teams so they can avoid similar pitfalls and also benefit from your successes…YOUR Team Experience Assessment report needs to totally (brutally) honest…but you might also want to create a ‘distilled’ version that you can share with other Data Mining teams and possibly management as a Lessons Learned or Process Improvement document…