ML Project Management Guidelines

Project Management

Data science projects traditionally involve a long exploration phase and many unknowns even quite late into a project. This is different from traditional software development, where it is possible to enumerate and quantify tasks at the outset.

We shall Divide our Project in below Phases

1. Ask
2. Prepare
3. Process
4. Analyse/Modelling
5. Evaluation and fine tuning
6. Production and Deployment
7. Share and Delivery
8. Act/Feedback

ASK Phase

In ASK phase we will try to ask guiding question related to project and try to identify its answers so that developing team and stakeholders both will be on same page with reference to project expectation and deliverable.

ASK Phase
Guiding questions • What is the problem you are trying to solve? • How can your insights/solution drive business decisions? • Is stated problem is the Root Cause?
Key tasks 1. Identify the business task 2. Consider key stakeholders
Deliverable A clear statement of the business task

Prepare Phase

Once we get proper understanding of business problem, we can think about how to solve it. Now it is Time to decide what data needs to be collected in order to answer the questions

In this phase we will **collect data, check its integrity and describe it**. We will also perform **data audit** to maintain record of data origin and its life cycle.

Following tasks will help us to prepare our data for further analysis

- Data Audit
- Describe data
- Explore data
- Verify data quality
- Data quality report

Data audit report

A Data Audit report consists of

1. Data Catalogue
2. Data Quality Scorecard

Data Catalogue

This catalogue documents the current understanding of the data source, communicates to stakeholders the sources to be used and some basic facts about them, and helps identify potential mismatches, concerns, or clarify misunderstandings.

Another simple method of identifying potential challenges is to clearly diagram the data pipeline used to build the model, showing where all data is from and how it is transformed.

Source	Contents	Duration	Quantity	Comments
Data source #1: data lake	Clickstream data	Jan 2018– Jan 2019	1.6M	User IP address only; user name not known
Data source #2: data lake	Order history	June 1 2016– Oct 3 2018	55k orders	Format stored in changed on Jan 1 2018 Final order only (not change history) Orders with errors are deleted
Sensor data	Readings from factory sensors. Streaming data is batched and stored	90 days history retention only	50/sec; 5k/sec expected	Data cleaning unknown; is perceived outlier data being dropped?

Data Quality Scorecard

The success of a Data project depends on the signal inherent in and extracted from the data. The Data Quality scorecard highlights areas that are frequently problematic in Data analysis and data science projects, and that can easily mislead the project—missing a signal that exists in the data or believing a signal exists where there is none—if not identified and addressed.

Example

Category	Example	Issue for project? (Y/N)	Status (Red/ Yellow/ Green)	Comments: Applicability, Status, Mitigations
Input data precision	Test & production data have same characteristics; outliers discarded for both model & production			
Input data accuracy	Sensor values estimated to be +- 5% of actual			
Data volumes & duration	Model data only available for 3 months (but business cycle is 1 year) 50% of source #3 data discarded			
Data sources & preprocessing validated	Data source #1 now undergoing additional quality checks Data extract #2 discovered to be flawed; re-training required			
Production vs model data pipeline	Prod inferences will use separate data source than model trained on			
Data change over time: processes considered	Upstream system changes logic & meaning of its input to model			

Note: This is example structure this may change depending upon type of data.

Describe data

Examine the "gross" or "surface" properties of the acquired data and report on the results.

Data description report

Describe the data that has been acquired including its format, its quantity (for example, the number of records and fields in each table), the identities of the fields and any other surface features which have been discovered. Evaluate whether the data acquired satisfies your requirements.

Explore data

During this stage you'll address data mining questions using querying, data visualization and reporting techniques. These may include:

- Distribution of key attributes (for example, the target attribute of a prediction task)
- Relationships between pairs or small numbers of attributes
- Results of simple aggregations
- Properties of significant sub-populations
- Simple statistical analyses

These analyses may directly address your data mining goals. They may also contribute to or refine the data description and quality reports, and feed into the transformation and other data preparation steps needed for further analysis.

Data exploration report

Describe results of your data exploration, including first findings or initial hypothesis and their impact on the remainder of the project. If appropriate you could include graphs and plots here to indicate data characteristics that suggest further examination of interesting data subsets.

Verify data quality

Examine the quality of the data, addressing questions such as:

- Is the data complete (does it cover all the cases required)?
- Is it correct, or does it contain errors and, if there are errors, how common are they?
- Are there missing values in the data? If so, how are they represented, where do they occur, and how common are they?

Data quality report

List the results of the data quality verification. If quality problems exist, suggest possible solutions. Solutions to data quality problems generally depend heavily on both data and business knowledge.

So to sum up, good data should be Original data from a Reliable organization, Comprehensive, Current, and Cited. **(ROCCC)**

Prepare Phase
Guiding questions • Where is your data located? • How is the data organized? • Are there issues with bias or credibility in this data? Does your data ROCCC? • How are you addressing licensing, privacy, security, and accessibility? • How did you verify the data's integrity? • How does it help you answer your question? • Are there any problems with the data?
Key tasks 1. Download data and store it appropriately. 2. Prepare audit report. 2. Identify how it's organized. 3. Sort and filter the data. 4. Determine the credibility of the data.
Deliverable A description of all data sources used and report.

Process Phase

Select your data

This is the stage of the project where you decide on the data that you're going to use for analysis. The criteria you might use to make this decision include the relevance of the data to your data mining goals, the quality of the data, and also technical constraints such as limits on data volume or data types. Note that data selection covers selection of attributes (columns) as well as selection of records (rows) in a table.

Rationale for inclusion/exclusion

List the data to be included/excluded and the reasons for these decisions.

Clean your data

This task involves raise the data quality to the level required by the analysis techniques that you've selected. This may involve selecting clean subsets of the data, the insertion of suitable defaults, or more ambitious techniques such as the estimation of missing data by modelling.

Data cleaning report

Describe what decisions and actions you took to address data quality problems. Consider any transformations of the data made for cleaning purposes and their possible impact on the analysis results.

The common information a Data cleaning report contains

- Major data cleaning operations performed
- Records dropped, either as an actual count or as a percentage
- Major issues found with the data, for example, "duplicate records found and dropped"
- Assumptions made at the time, for example, "data was extracted for US only, other countries are assumed to be similar"

Construct required data

This task includes constructive data preparation operations such as the production of derived attributes or entire new records, or transformed values for existing attributes.

- **Derived attributes** – These are new attributes that are constructed from one or more existing attributes in the same record, for example you might use the variables of length and width to calculate a new variable of area.
- **Generated records** – Here you describe the creation of any completely new records. For example you might need to create records for customers who made no purchase during the past year. There was no reason to have such records in the raw data, but for modelling purposes it might make sense to explicitly represent the fact that particular customers made zero purchases.

Integrate data

These are methods whereby information is combined from multiple databases, tables or records to create new records or values.

Merged data

Merging tables refers to joining together two or more tables that have different information about the same objects.

For example a retail chain might have one table with information about each store's general characteristics (e.g., floor space, type of mall), another table with summarised sales data (e.g., profit, percent change in sales from previous year), and another with information about the demographics of the surrounding area. Each of these tables contains one record for each store. These tables can be merged together into a new table with one record for each store, combining fields from the source tables.

Aggregations

Aggregations refers to operations in which new values are computed by summarising information from multiple records and/or tables.

For example, converting a table of customer purchases where there is one record for each purchase into a new table where there is one record for each customer, with fields such as number of purchases, average purchase amount, percent of orders charged to credit card, percent of items under promotion etc.

Process Phase
Guiding questions • What tools are you choosing and why? • Have you ensured your data's integrity? • What steps have you taken to ensure that your data is clean? • How can you verify that your data is clean and ready to analyze? • Have you documented your cleaning process so you can review and share those results?
Key tasks 1. Check the data for errors. 2. Choose your tools. 3. Transform the data so you can work with it effectively. 4. Document the cleaning process.
Deliverable Documentation of any cleaning or manipulation of data

Analyse/Modelling

Next up is to make some conclusions based on the trustable data. **Data Analyses** is a skill that takes time to master, but over time the patterns will emerge faster and methods one uses will develop. Main concept is to think analytically about your data, be critical and be creative.

There might be a need to sort and format the data to make it easier to process, make a Pivot table, or create awesome graphs! Remember it is a story that must unfold. Further processing might include:

- Performing different calculations get get additional metrics.
- Combining additional data attributes from a variety of sources to get a more comprehensive story.
- Create different views for the data. Like tables with your results, filter and pivot them.
- Make it visual if possible! Charts tell more than a thousand words.

Modelling

Select modelling technique

As the first step in modelling, you'll select the actual modelling technique that you'll be using. Although you may have already selected a tool during the business understanding phase, at this stage you'll be selecting the specific modelling technique e.g. decision-tree building with C5.0, or neural network generation with back propagation. If multiple techniques are applied, perform this task separately for each technique.

- **Modelling technique** – Document the actual modelling technique that is to be used.
- **Modelling assumptions** – Many modelling techniques make specific assumptions about the data, for example that all attributes have uniform distributions, no missing values allowed, class attribute must be symbolic etc. Record any assumptions made.

Generate test design

Before you actually build a model you need to generate a procedure or mechanism to test the model's quality and validity. For example, in supervised data mining tasks such as classification, it is common to use error rates as quality measures for data mining models. Therefore, you typically separate the dataset into train and test sets, build the model on the train set, and estimate its quality on the separate test set.

- **Test design** – Describe the intended plan for training, testing, and evaluating the models. A primary component of the plan is determining how to divide the available dataset into training, test and validation datasets.

Build model

Run the modelling tool on the prepared dataset to create one or more models.

- **Parameter settings** – With any modelling tool there are often a large number of parameters that can be adjusted. List the parameters and their chosen values, along with the rationale for the choice of parameter settings.
- **Models** – These are the actual models produced by the modelling tool, not a report on the models.
- **Model descriptions** – Describe the resulting models, report on the interpretation of the models and document any difficulties encountered with their meanings.

Assess model

Interpret the models according to your domain knowledge, your data mining success criteria and your desired test design. Judge the success of the application of modelling and discovery techniques technically, then contact business analysts and domain experts later in order to discuss the data mining results in the business context. This task only considers models, whereas the evaluation phase also takes into account all other results that were produced in the course of the project.

At this stage you should rank the models and assess them according to the evaluation criteria. You should take the business objectives and business success criteria into account as far as you can here. In most data mining projects a single technique is applied more than once and data mining results are generated with several different techniques.

- **Model assessment** – Summarise the results of this task, list the qualities of your generated models (e.g.in terms of accuracy) and rank their quality in relation to each other.
- **Revised parameter settings** – According to the model assessment, revise parameter settings and tune them for the next modelling run. Iterate model building and assessment until you strongly believe that you have found the best model(s). Document all such revisions and assessments.

Analyse/Modelling Phase
Guiding questions • How should you organize your data to perform analysis on it? • Has your data been properly formatted? • What surprises did you discover in the data? • What trends or relationships did you find in the data? • How will these insights help answer your business questions? • Which features should be selected? • Which is the best model as per available data?
Key tasks 1. Aggregate your data so it's useful and accessible. 2. Organize and format your data. 3. Perform calculations, Modelling and feature selection 4. Identify trends and relationships.
Deliverable A summary of your analysis and Modelling

Evaluation/Fine Tune

Evaluate your results

Evaluation steps deals with factors such as the accuracy and generality of the model. During this step you'll assesses the degree to which the model meets your business objectives and seek to determine if there is some business reason why this model is deficient.

Another option is to test the model(s) on test applications in the real application, if time and budget constraints permit. The evaluation phase also involves assessing any other data mining results you've generated.

Data analysis results involve models that are necessarily related to the original business objectives and all other findings that are not necessarily related to the original business objectives, but might also unveil additional challenges, information, or hints for future directions.

Assessment of data analysis results

Summarise assessment results in terms of business success criteria, including a final statement regarding whether the project already meets the initial business objectives.

Approved models

After assessing models with respect to business success criteria, the generated models that meet the selected criteria become the approved models.

Review process

At this point, the resulting models appear to be satisfactory and to satisfy business needs. It is now appropriate for you to do a more thorough review of the process engagement in order to determine if there is any important factor or task that has somehow been overlooked.

This review also covers quality assurance issues—for example: did we correctly build the model? Did we use only the attributes that we are allowed to use and that are available for future analyses?

Review of process

Summarise the process review and highlight activities that have been missed and those that should be repeated.

Determine next steps for fine tuning

Depending on the results of the assessment and the process review, you now decide how to proceed. Do you finish this project and move on to deployment, initiate further iterations, or set up new data mining projects? You should also take stock of your remaining resources and budget as this may influence your decisions.

List of possible actions

List the potential further actions, along with the reasons for and against each option.

Evaluation and Fine tune Phase
Guiding questions • How good is model accuracy? • Is resultant output meets business objective? • How our model output affects business KPIs? • Is there any fine tuning is required?
Key tasks 1. List further corrective action to improve model 2. Evaluation of model in business operation 3. Improve model performance 4. Review process to highlight activities that have been missed and those that should be repeated.
Deliverable High performing model which can resolve business task

Production & Deployment

Plan deployment

In the deployment stage you'll take your evaluation results and determine a strategy for their deployment.

If a general procedure has been identified to create the relevant model(s), this procedure is documented here for later deployment. It makes sense to consider the ways and means of deployment during the business understanding phase as well, because deployment is absolutely crucial to the success of the project. This is where predictive analytics really helps to improve the operational side of your business.

Deployment plan

Summarise your deployment strategy including the necessary steps and how to perform them.

Plan monitoring and maintenance

Monitoring and maintenance are important issues if the modelling result becomes part of the day-to-day business and its environment. The careful preparation of a maintenance strategy helps to avoid unnecessarily long periods of incorrect usage of data mining results. In order to monitor the deployment of the data analysis result(s), the project needs a detailed monitoring process plan. This plan takes into account the specific type of deployment.

Monitoring and maintenance plan

Summarise the monitoring and maintenance strategy, including the necessary steps and how to perform them.

Production & deployment Phase
Guiding questions • What are the best method for deploying model? • Is product and deployment method satisfying client need? • is our deployment compromising model operation?
Key tasks 1. Validate deployment operation 2. Ensure effective working of model 3. Document production steps
Deliverable Optimum and effective deployment method

Share and delivery

Produce final report and delivery document

At the end of the project you will write up a final report. Depending on the deployment plan, this report may be only a summary of the project and its experiences (if they have not already been documented as an ongoing activity) or it may be a final and comprehensive presentation of the data analysis result(s).

- **Final report** – This is the final written report of the data mining engagement. It includes all of the previous deliverables, summarising and organising the results.
- **Final presentation** – There will also often be a meeting at the conclusion of the project at which the results are presented to the customer.
- **Delivery Document**- Should include goals and final outcome with mentioning key stack holders. Method/Guideline for implementation of product.

Share and delivery Phase
Guiding questions • What story does your data tell? • How do your findings relate to your original question? • Who is your audience? What is the best way to communicate with them? • Can data visualization help you share your findings? • Is your presentation accessible to your audience?
Key tasks 1. Determine the best way to share your findings. 2. Create effective data visualizations. 3. Present your findings. 4. Ensure your work is accessible.
Deliverable Key findings and product implementation

Act/Feedback

Review project

Assess what went right and what went wrong, what was done well and what needs to be improved.

Experience documentation

Summarise important experience gained during the project. For example, any pitfalls you encountered, misleading approaches, or hints for selecting the best suited data mining techniques in similar situations could be part of this documentation.

Also, it is very important to get clients feed back once you achieve the business goal. This type of feedback can be taken from client in form of appreciation document or testimonials.

Evaluation and Fine tune Phase
Guiding questions • What are major challenges faced during project? • What are the client's feedback?
Key tasks 1. Get in touch with end user to take their feedback 2. Identify different challenges faced during project.
Deliverable Client testimonial and signed document