Data Science Project Checklist

Created by Jake Strasler

Last updated: January 10, 2024

This checklist is a print-friendly, interactable translation of DataCamp's Data Science Project Checklist. DataCamp describes the checklist as "summarizing data science project management best practices..." and "...combines CRISP-DM project management principles with those of the Agile and Scrum software development frameworks."

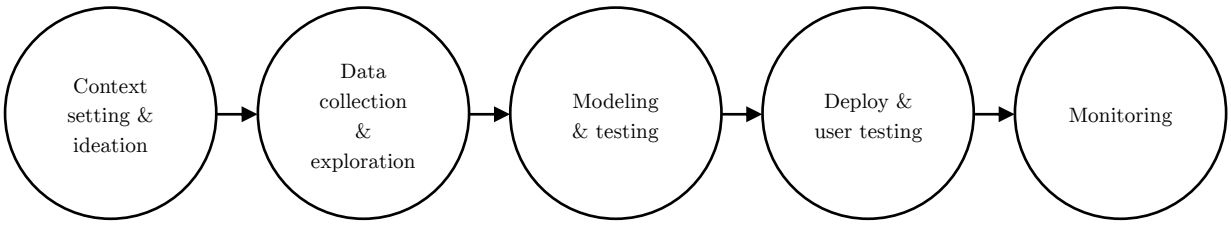
The following characteristics should underpin any data science project:

- **Measurable**: Is the success of a project and its impact on the business quantifiable?
- **Reliable**: What proportion of projects achieved their goals?
- **Scalable**: Can the throughput of projects be increase without significantly degrading reliability?

Additionally, the following principles should be abided by:

- **Iterative**
- **Reusable and recyclable**
- **Reproducible**

The framework followed will be:



Project Name: _____

Project Start Date: _____

Notes: _____

_____

Context Setting & Ideation

	Task	Notes
☐	Identify the business problem being solved. *Clarify why the project is being undertaken as unambiguously as possible.*	
☐	Identify stakeholders. *Roles may include project manager, data scientist, account manager, etc.*	
☐	Review prior work. *Review existing projects that covered similar ground. What were key outcomes? Can anything be reused? What mistakes should be avoided?*	
☐	Determine key performance indicators to measure success. *Metrics should use SMART criteria.*	
☐	Determine the scope. *What are the deliverables? What are the reqs. for the deliverables? What will not be included?*	
☐	Write a project plan. *Set milestones for intermediate steps in the project. Decide on timeline for each milestone. Write a short description of each step.*	Do external to this document.

☐	Estimate the impact of the project. *Quantify the benefit to the organization if the goals are reached. List qualitative benefits.*	
☐	Estimate the effort of the project. *How much will the project cost? How much time will the project take? What resources will it need?*	
☐	Estimate the project risks. *List all the risks to the project. Calculate the risk impact the probability of it occurring times severity of it occurring.*	
☐	Decide whether or not to proceed with the project. *Decide to proceed now, put on hold, or cancel.*	
☐	Determine the responsibility of each stakeholder. *Use the RACI model for each task.*	
☐	Determine a communication strategy. *How will you keep in touch? Meeting cadence?*	
☐	Identify data sources. *Do you have access? Where is it stored? What form is it in? Etc.*	
☐	Anticipate regulatory needs.	

	Will anything be audited? Can data sources be legally used?	
☐	Decide on a technology stack. *Agree on tools for storing, processing, and modeling the data.*	
☐	Write a project charter. *Summarize what was decided in a short document.*	Do external to this document.

Data Collection & Exploration

	Task	Notes
☐	Give data scientists access to all datasets. *Organize appropriate permissions. Purchase any commercial datasets or use synthetic data with similar properties.*	
☐	Ingest the data. *For each data source, move it to the analytics environment.*	
☐	Explore the data. *Visualize distribution of each variable with histogram or bar plot. Quantify missing values for each variable. Visualize relationship between features and target variable with scatter, histogram, box, or heatmap plots.*	
☐	Determine key performance indicators to measure success. *For each dataset: provide a summary, describe high-level data quality issues, describe the quality of target variable, describe quality of each feature, describe relationship between each feature and the target variable.*	

☐	Decide whether or not to proceed with the project. *Continue, pause, or cancel.*	
☐	Build a data pipeline. *Pipeline should automate the ingestion and cleaning process. Should run as batch or stream job.*	
☐	Document the data pipeline. *Draw diagram of steps in the data pipeline and their dependences. Describe what happens in each step.*	Do external to this document.

Modeling & Testing

	Task	Notes
	Modeling	
☐	Generate a hypothesis. *Does the hypothesis make sense in the business domain? Can you measure the outcome? Do you have sufficient data to be significant? Watch for biases.*	
☐	Split your data into training and testing sets. *Do this before engineering features to ensure no data leakage.*	
☐	Engineer features. *Center/scale numeric values, create categorical variables (incl. binning), apply Box-Cox or Yeo-Johnston transformations to numeric variables to give normal dist., combine rare/related categories of categorical variables, extract or combine parts of datetimes, create new variables from summary statistics, extract quantitative metrics from text and other unstructured data.*	
☐	Fit the model/run an experiment. *Fit simplest first then increase complexity.*	

☐	Report on the results. *Regularly provide feedback to stakeholders, adjust language accordingly, report failures as well as successes.*	
	Testing	
☐	Create a test suite. *Unit tests or back test that run automatically to check your model or experiment's performance.*	
☐	Validate the business impact. *Discuss the impact with business stakeholders.*	
☐	Validate the technical approach. *Check assumptions of your model to ensure they're valid. See if results are sensitive to sampled data. Are hyperparameters suitable? Is the model reproducible?*	
☐	Validate the deployability. *Can all input values or use cases be handled? Are all required data sources available in production? Can model fail gracefully? Can predictions be made fast enough?*	
☐	Preserve null results. *Record anything that won't make it to production.*	

Deployment & User Testing

	Task	Notes
	Deployment	
☐	Develop a data pipeline. *Set up DAG for all the data sources to production environment.*	
☐	Develop a model pipeline. *Provide an API to your model that can be accessed by dashboard/website/etc.*	
☐	Design a monitoring plan. *Determine metrics to be tracked, including performance and safety metrics. Determine limits for acceptable ranges for those metrics. Decide how you wish to be alerted if metrics go out of range.*	
☐	Roll out via A/B test. *Provide the new feature/model to random sample of users. Monitor chosen metrics. Declare winner when there is statistical significance.*	
☐	Analyze and report on A/B test results. *Compare the metrics chosen to track for each group. Report results even if test was not a success.*	
☐	Roll out to most or all users. *If successful, rollout to all but small holdout group.*	

User Testing	
☐ Write an exit report. *Summarize status of project and what you learned. Provide overview of the project. Summarize business problem you tried to solve. Describe the data sources and how they were processed. Describe modeling techniques used and their validation. Summarize solution architecture. Outline benefits from the project for company and customer. Describe any learnings around project execution, data science, business domain, and the product. Outline next steps.*	Do external to this document.
☐ Get customer feedback. *Conduct surveys and user interviews. Monitor reviews, ratings, and social media.*	

Monitoring

	Task	Notes
☐	Develop monitoring pipeline. *Set up pipeline to automatically track the performance and safety metrics defined in the monitoring plan.*	
☐	Create dashboards. *Create dashboards to track the changes of metrics over time.*	
☐	Set up alerts. *Set up notifications via email, Slack, etc. when metrics fall out of the acceptable range.*	

Project End Date: _____

Time Elapsed: _____