Decision Data

Data Science Project Structure

The CRISP-DM Methodology

#	Stage	Considerations
1	**Business understanding**	• **Problem statement:** Understand the context of the problem • **Objectives:** What is the purpose of the project? • **Needs:** What is the use case of the solution? Who are the users? Which features are necessary, and which are just nice to have? • **Data:** Identify data sources, an estimate of the data size and velocity, and data collection method. • **Definitions of success:** Determine the satisficing and optimising metrics, as well as the success and exit criteria. • **Presentation:** Dashboard, application, or report? Cloud or on-prem? • **Timelines:** Milestones
2	**Data understanding** *Return to stage 1 if required*	• **Data quality:** What are the quality issues and treatment options? • **EDA:** Inspect the distributions, statistical summaries, and units/scale. Explore whether there are any correlations or underlying patterns within the data.
3	**Data preparation**	• **Pre-processing:** Select and apply treatment to your data quality issues, do some feature engineering. • **ETL:** Integrate the data, perform transformations and scaling.
4	**Modelling** *Return to stage 3 if required*	• **Modelling**: Build a baseline model quickly, then iterate. • **Split:** Determine the train-dev-test split or consider cross-validation. • **Tuning:** Use the principles of orthogonalisation to guide you with hyperparameter tuning; consider grid search. Apply optimisation techniques if required. • **Comparison:** Use your metrics to guide whether your current idea is better than your previous idea. • **Iterate:** Iterate until either the success or exit criteria have been met.
5	**Evaluation** *Return to stage 1 if required*	• **Write-up:** Document the results of your trials as well as any reasoning for your decisions. • **Evaluation**: Ask the stakeholder to assess your final model against the business requirements.
6	**Deployment**	• **Plan:** Create a deployment plan and checklist. • **Clean up:** Convert any notebooks into scripts, draw diagrams for each script. • **Deploy:** Deploy your project, provide access to the stakeholders, and monitor the results.