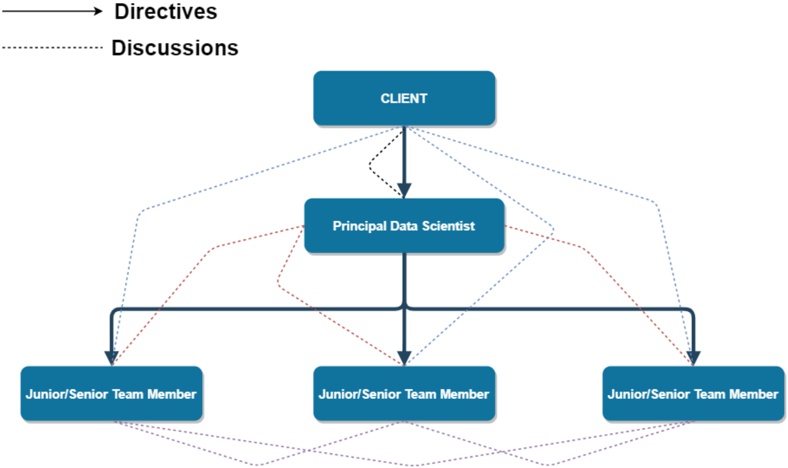
Project Management and Data Science

Ryan Miller

Organizational Structure

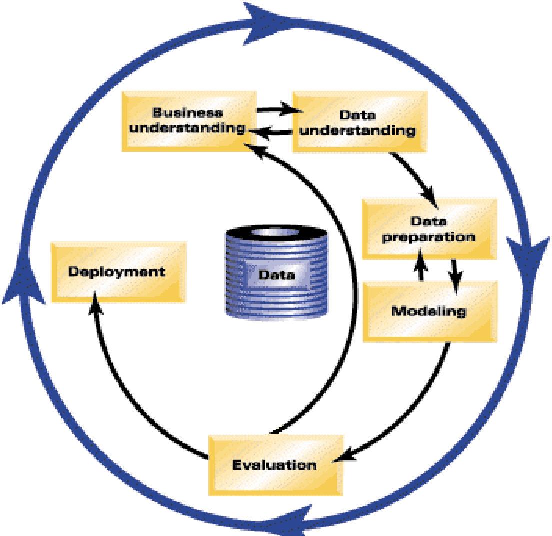

Directives
Discussions
CLIENT
Principal Data Scientist
Junior/Senior Team Member
Junior/Senior Team Member
Junior/Senior Team Member

Data Science “pipelines” and “life-cycles”

- The data science work-flow takes place within the organizational framework
- While “Data science” is oft used, its definitions vary
  - Click here to see some examples of this

CRISP-DM

We'll adopt the Cross-Industry Standard Process for Data Mining (CRISP-DM) model:



Phase 1 - Business Understanding

Tasks:

1. Gather necessary background information
2. Document specific specific objectives
3. Determine success criteria for the project

Each of these tasks should be undertaken in coordination with the client

Phase 1 - Objective vs. Subjective Success Criteria

- You may choose to have a mixture of objective and subjective success criteria
  - Objective = “Increase the time visiters spend on the landing page by 10%”
  - Subjective = “Identify customer clusters for targeted marketing”

Phase 2 - Data Understanding

Tasks:

1. Describing the data
2. Exploring the data
3. Verifying data quality

These tasks should be carried out at the team level (and cross-referenced with the principal and client if necessary)

Phase 3 - Data Preparation

Tasks:

- Merging/joining (ie: `left_join`)
- Selecting relevant subsets (ie: `filter`)
- Aggregating records (ie: `group_by` and `summarize`)
- Deriving new attributes (ie: `mutate`)
- Handling missing data (ie: `complete.cases` or `knnImput`/`rfImpute`)

These tasks should be carried out at the team level and cross-referenced by the principal (they are seldom relevant to the client at this point)

Phase 4 - Modeling

Tasks:

- Selecting a model
- Evaluating the “goodness” of a model
- Building the model
- Note: you may replace “model” with “product” in some applications

This phase is highly non-linear, it should be carried out at the team level and cross-referenced by the principal (and possibly with the client depending on their level of technical proficiency)

Phase 5 - Evaluation

Tasks:

1. Consider your model/product in regards to the business success criteria you came up with in Phase 1
2. Formalize your findings

These tasks should be undertaken in coordination with the client and principal

Phase 6 - Deployment

Tasks:

1. Deliver your model/product to the client
2. Complete wrap-up tasks (ie: technical report, etc.)

Practice #1

For the following scenario determine:

1. Which phase the described actions fall under
2. Where you'd go next (and why)

> *A project is using medical records to build a model to predict A1c levels using more readily available measures such as blood pressure, age, weight, and waist circumference. Using the `is.na` function in `R`, it is discovered that 88% of the available records do not have an A1c measurement.*

Practice #1 - Possible Answers

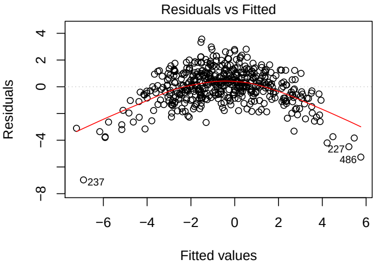
- This scenario occured during the Data Understanding phase
- It is possible that the 12% of medical records with A1c provides a reasonable dataset, in this case the next step would be Data Preparation (filtering out the missing data and preparing the other variables)
- It is also possible that relying on only 12% of the available records is infeasible or will induce bias into the analysis, in this case the next step would be Business Understanding (reviewing the original goals and making adjustments)

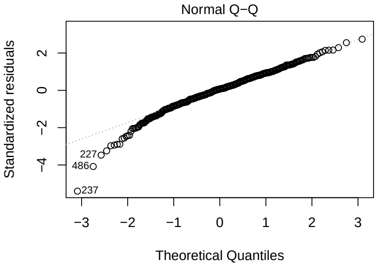
Practice #2

For the following scenario determine:

1. Which phase the described actions fall under
2. Where you'd go next (and why)

In the aforementioned project you fit a linear regression model containing several variables, you receive the following model diagnostics from your software





Practice #2 - Possible Answers

- This scenario occured during the Modeling phase
- Linear regression doesn't appear to be an appropriate model based upon these diagnostics. It seems that a quadratic effect is being missed (hence the large negative residuals for high/low fitted values). The QQ-plot also calls into question whether the residuals are normally distributed (not a disaster for model fitting, but a problem for statistical inference)
    - For these reasons the logical next step is to return to the Data Prepartion phase and explore variable transformations that might address these issues

Practice #3

For the following scenario determine:

1. Which phase the described actions fall under
2. Where you'd go next (and why)

After revising the linear regression model in the previous example, a final model is chosen and is applied a "test set" of 100 new records that occured after the original dataset was finalized. The model predicts A1c within 10% of the actual value for 86% of these new records.

Practice #3 - Possible Answers

- This scenario occured in the Evaluation phase
- Where to go next *depends upon the project's business goals*. If predicting A1c within 10% of the actual value for 86% of cases satisfies the previously established goals, the model is suitable for the Deployment phase. Otherwise the project might need to return to square one.