Data Science Project Workflow

(1) Preparation

Base Questions:

- What are we trying to build/do?
- Who is responsible for what task?
- What are the deadlines?
- What is the budget?
- Who is the target group?

Project Questions:

- Exploration, Regression, Classification, Hypothesis Test?
- Continuous calculation / maintenance?
- Hardware availability?
- Target metrics?

(2) Data Acquisition

- 1.A. Own Database
  - SQL, CSV
- 1.B. External Database (Web, other Companies)
  - SQL, WEBCRAWLING, CSV
- 1.C. Field Research
  - Actual field research, Web statistics

(2.5) Data Pipeline

- Automate Data Query
  - SQL, PYTHON/R, EXCEL, COMMANDLINE, SPARK, HADOOP → IT-Department

(3) Data Transformation

- First Glimpse at Data
  - Missing variables?
- Deal with NAs
- Create working sample, if data is large
- Check for extreme values
- Feature Engineering
- For Visualization:
  - Grouping
  - Scaling
- For Modeling:
  - Train/Test-Split
  - OH-Encoding
  - Scaling

(4) Explorative Analysis

- Numeric Analysis
  - Ranges
  - Missing Values
  - Variance
  - Correlations (Correlation Matrix)
- Visual Analysis
  - Distributions (Histograms)
  - Differences between Groups (Boxplots)
  - Correlations (Scatterplots)
- Feature (Re-)Engineering

(5) Modeling

Some Options:

- 5.A. Linear/Logistic Regression:
  - Shows variable impact on model (coefficients)
  - Usually underperforms other models in Prediction/Classification
  - Easy to understand
  - Can use weights
  - Hypothesis tests possible
- 5.B. Random Forest:
  - Good baseline for prediction/classification
  - Shows feature importance
  - Grid search to tune hyperparameters
- 5.C. XGBoost:
  - Boosted models can outperform Random Forests
  - Grid Search to tune Hyperparameters
  - Black-Box method
- 5.D. Artificial Neural Network (ANN):
  - Works well on large datasets
  - Best for human-like learning (e.g. image recognition)
- 5.E. Stacking:
  - Stack predictions from multiple models

(6) Production/Results

Always show process: What have we done to come to this result? (short, adequate for target group)

- 6.A. Deliver Insight:
  - Visualization
  - Business Action
- 6.B. Hypothesis Test:
  - Could the H_0 be rejected?
- 6.C. Deliver Predictions or Classification
  - Visualization
  - Business Action

(6.5) Production Pipeline

- Automate Prediction/Classification
  - SQL, PYTHON/R, EXCEL, COMMANDLINE, SPARK, HADOOP → IT-Department
- Dashboarding
  - HTML, CSS, JavaScript, EXCEL
- Large Datasets:
  - Implement model on suited Hardware

GITHUB.COM/CHOLZKORN