

PRACTICE DAY

Naumaan Nayyar

OBJECTIVES

- ▶ Practice the data science workflow on problems

REVIEW

THE DATA SCIENCE WORKFLOW

OVERVIEW OF THE DATA SCIENCE WORKFLOW

The steps:

1. Identify the problem
2. Acquire the data
3. Parse the data
4. Mine the data
5. Refine the data
6. Build a data model
7. Present the results



OVERVIEW OF THE DATA SCIENCE WORKFLOW



IDENTIFY THE PROBLEM

- ☐ Identify business/product objectives
- ☐ Identify and hypothesize goals and criteria for success
- ☐ Create a set of questions for identifying correct data set

OVERVIEW OF THE DATA SCIENCE WORKFLOW



ACQUIRE THE DATA

- ☐ Identify the “right” data set(s)
- ☐ Import data and set up local or remote data structure
- ☐ Determine most appropriate tools to work with data

OVERVIEW OF THE DATA SCIENCE WORKFLOW

PARSE THE DATA

- ☐ Read any documentation provided with the data
- ☐ Perform exploratory data analysis
- ☐ Verify the quality of the data

OVERVIEW OF THE DATA SCIENCE WORKFLOW



MINE THE DATA

- ☐ Determine sampling methodology and sample data
- ☐ Format, clean, slice, and combine data in Python
- ☐ Create necessary derived columns from the data (new data)

OVERVIEW OF THE DATA SCIENCE WORKFLOW



REFINE THE DATA

- ☐ Identify trends and outliers
- ☐ Apply descriptive and inferential statistics
- ☐ Document and transform data

OVERVIEW OF THE DATA SCIENCE WORKFLOW



BUILD A DATA MODEL

- ☐ Select appropriate model
- ☐ Build model
- ☐ Evaluate and refine model

1. Data Collection

2. Data Cleaning

3. Data Exploration

4. Feature Engineering

5. Model Selection

6. Model Training

7. Model Evaluation

8. Model Deployment

9. Monitoring

10. Retention

11. Feedback

12. Iteration

13. Reporting

14. Archiving

15. Documentation

OVERVIEW OF THE DATA SCIENCE WORKFLOW



PRESENT THE RESULTS

- ☐ Summarize findings with narrative, storytelling techniques
- ☐ Present limitations and assumptions of your analysis
- ☐ Identify follow up problems and questions for future analysis

GROUP ACTIVITY

PRACTICE

ACTIVITY: DATA SCIENCE PRACTICE

DIRECTIONS (120 minutes)



EXERCISE

1. Form groups.
2. Pick a problem among: loan funding prediction, sentiment analysis, housing prices prediction.
3. Go through the data science workflow.