AN ABSTRACTED PYTHON CLASS FOR TDSP ML PROJECTS: EXAMPLE USE CASE WITH WINE ID

Kevin Geidel

MSDS 422: Practical Machine Learning

Northwestern University

June 2, 2024

1 Executive summary

Given the rising preveleance of data science teams in industry and the widespread adoption of the *Team Data Science Process (TDSP)* methodology, an abstracted library that streamlines common TDSP projects would allow "citizen developers" in data related roles to increase the efficiency and performance of their *Machine Learning (ML)* applications. There are commerical data science solutions available for purchase. However, many small teams are using agile project management, cloud computing resources and versitile opensource frameworks to create low cost, high performaning analytics and data science tools that are tailor made for many different settings throughout various enterprises (Hyatt, 2024).

The proposed `Python` classes create a work flow that mirrors the TDSP (illustrated in figure 1.) The benefits of using customized libraries to conduct ML projects are shortened development cycles, more robust/stable applications, greater interoperability amongst team members and more time to focus on making models that perform better and provide more actionable conclusions.

2 Research design

-Can we abstract datamining to automate model evaluation and selection? -In a big MS thesis like data sync design where you use ORM youll want a custom pipeline toolkit -Wine defect prediction use case

3 Exploratory Data Analysis (EDA)

4 Datamining data models

5 Data pipeline and ML preprocessing

6 ML engineering, evaluation and deployment

- data transformations!

7 Findings

8 Conclusions

Appendix

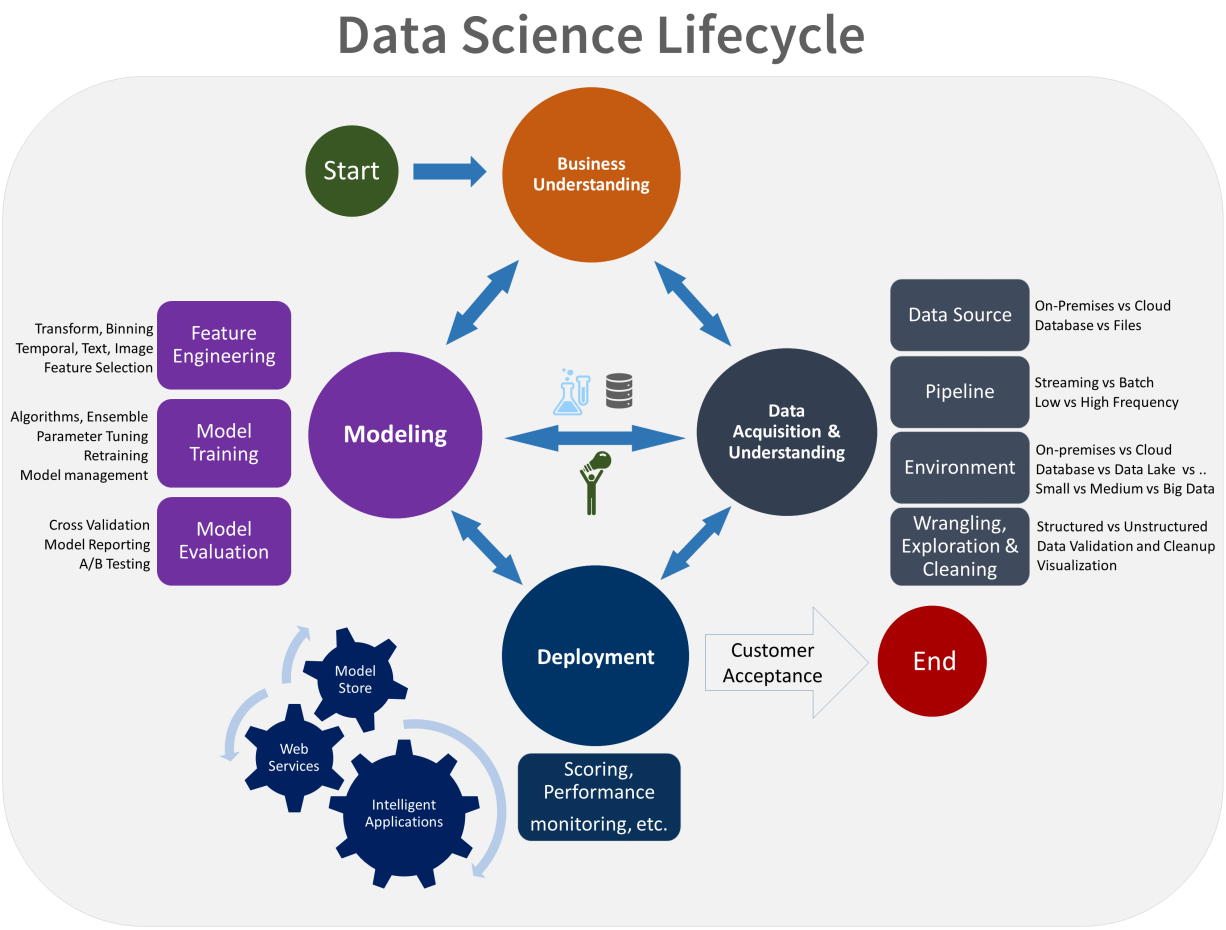

Figure 1: (color online) The "data lifecycle" in Team Data Science Process (TDSP)

References

Hyatt, J. (2024). Django + postgres: A data science juggernaut. Master's thesis, Syracuse University.