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 ("What is the Team Data Science Process?", 2024). Various stages in the TDSP (which is illustrated in figure 1) are represented by data models and have methods for applying standard work to these structures, moving them along the path to the next stage. 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. The code for the project (and this paper) can be found in the `GitHub` repositiory located at `https://github.com/kgeidel/MSDS-422-final-project` (Geidel, 2024).

The concept is explored with a use case that seeks to deploy a binary classifier for a defect prediction application. Using the `wine` dataset (Vanschoren, 2014) from `openml.org` we have 14 features for 178 records that include labels on if the sameple is wine or not. *Exploratory Data Analysis (EDA)* is conducted with the aim of selecting, training and tuning four models to classify, for now, a sample as wine or not. This would be extended to a model that seeks to predict samples that run a high probability of being defective or non-conforming. Emphasis is on turning this wine classifier deployment into a DRY (*Don't Repeat Yourself*), generic (but still flexible) and easy to deploy tool that data scientists can use to expedidate routine data exploration, data wrangling and datamining.

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

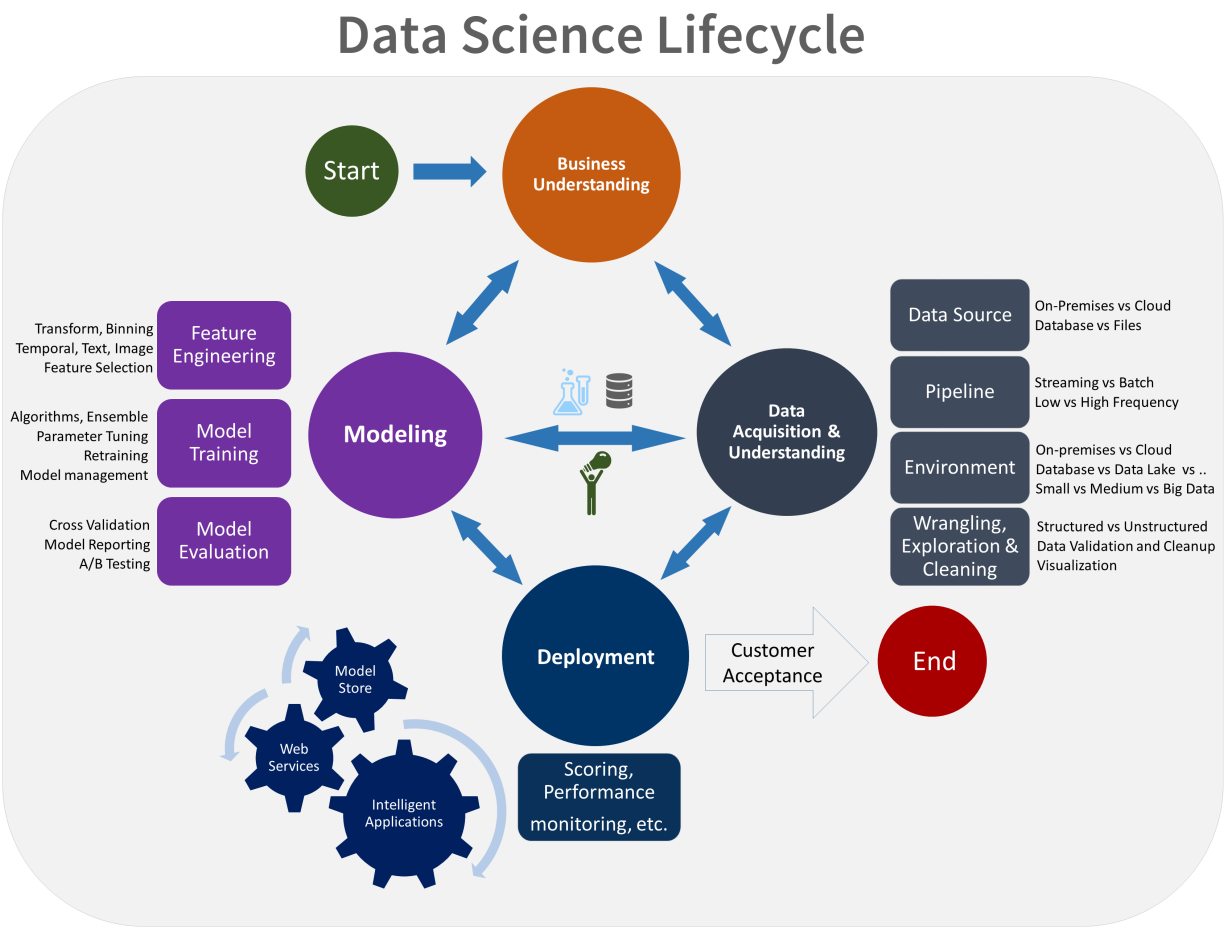

Figure 1: (color online) The "data lifecycle" in Team Data Science Process (TDSP)

References

Geidel, K. (2024). MSDS-422-final-project. Northwestern University. https://github.com/kgeidel/MSDS-422-final-project.

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

Vanschoren, J. (2014). wine. https://www.openml.org/search?type=data&sort=runs&status=active&id=973.

"What is the Team Data Science Process?" (2024). Azure Architecture Center. https://learn.microsoft.com/en-us/azure/architecture/data-science-process/overview.