DVC and Kedro

2022 • GIORGIA BERTACCHINI

DVC (Data Version Control) and kedro

DATA VERSION CONTROL

What is

- It takes on a Git-like model to provide management and versioning of datasets and machine learning models. DVC is a simple command-line tool that makes machine learning projects shareable and reproducible.

DATA

data/data.xml.dvc

- DVC stores information about the added file in a special .dvc file named data/data.xml.dvc, this metadata file is a placeholder for the original data.

KEDRO

What is

- Kedro is an open-source Python framework for creating reproducible, maintainable and modular data science code.

DATA

conf/base/catalog.yml

- Data Catalog, which is the registry of all data sources available for use by the project.

directory /data

- where the data are divided during project.
- where are saved also models, plot and other created.

DVC (Data Version Control) and Kedro

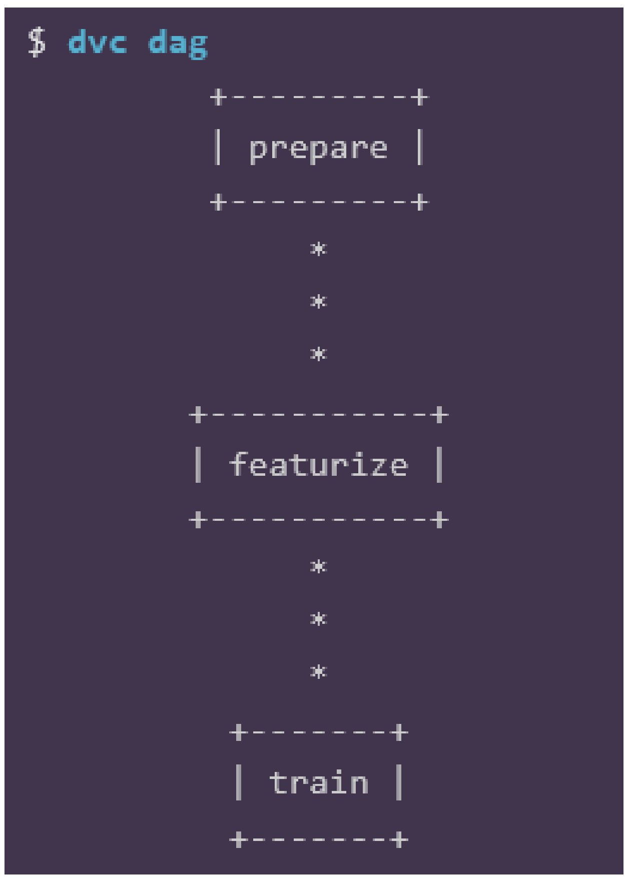
PIPELINE

dvc.yaml file

- It includes information about the steps of pipeline, with dependencies and outputs, and concatenate the nodes od pipeline

command: dvc dag

- to visualize the pipeline structure.



```
$ dvc dag
          +---------+
          | prepare |
          +---------+
               *
               *
               *
         +-----------+
         | featurize |
         +-----------+
               *
               *
               *
           +-------+
           | train |
           +-------+
```

PIPELINE

src/name_kedro_project/pipelines/name_pipeline/pipeline.py file

- It includes information about the steps of pipeline, with dependencies and outputs, and concatenate the nodes od pipeline

command kedro viz

- this command should open up a visualisation in your browser
- to visualize the pipeline structure and other informations

DVC (Data Version Control) and Kedro

METRICS

DVC makes it easy to track metrics, and visualize performance with plots.

command: dvc run

- specifing node of pipeline and dependencies, create in output plots and a file with metrics.

command show diff

- show difference through metrics different, for example metrics of different branches

EXPERIMENTS

DVC can track the experiments, list and compare their most relevant metrics, parameters.

METRICS

command kedro viz

- to visualize same data, for example MetricsDataSet and PlotlyDataSet and other informations

EXPERIMENTS

command kedro viz

- Experiment tracking in Kedro-Viz also supports the display of plots, such as Plotly and Matplotlib, and other results from all experiments.

DVC (Data Version Control) and Kedro

PARAMETERS

params.yaml

- DVC can track parameters, that can be any values used inside your code to influence the results.

command: dvc params diff

- Show changes in dvc params between commits in the DVC repository

PLOTS

DVC have a set of commands to create, visualize and compare data sets.

PARAMETERS

parameters/name_pipeline.yaml

- where are write parameters, that can be any values used inside your code to influence the results.

PLOTS

command kedro viz

- Kedro-Viz show the plot of data in output of PlotlyDataSet and Matplotlib nodes.

DVC (Data Version Control) and Kedro

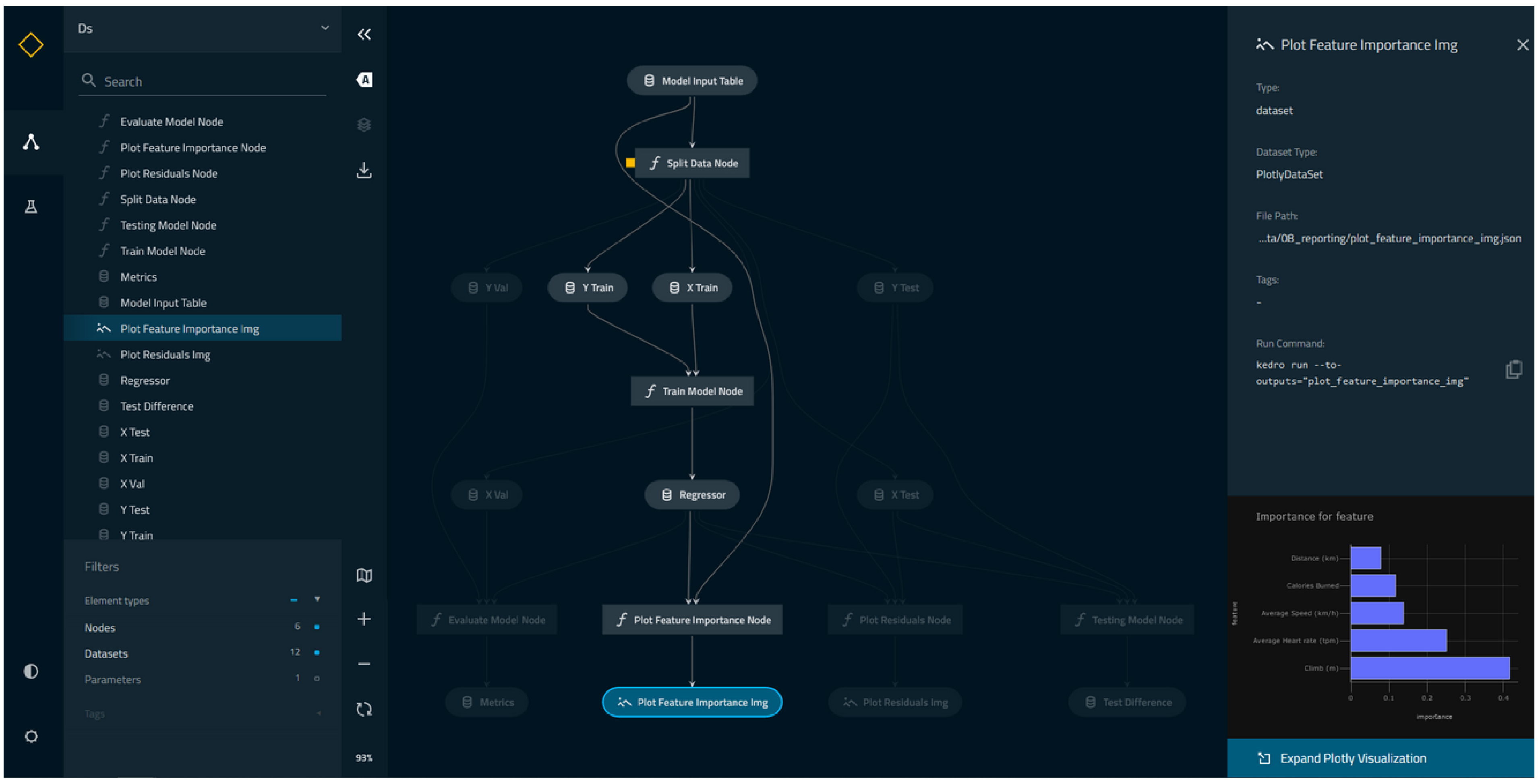
DVC ON VS CODE

There are a DVC extension, which brings a full machine learning experimentation platform to Visual Studio Code.
with this extension in VisualCode can have Interactive plots, Live tracking and Experiment bookkeeping.

More: https://iterative.ai/blog/DVC-VS-Code-extension

KEDRO-VIZ

The same feature of DVC extension for Visual Studio Code are also in the browser opened by command "kedro viz".



Plot Feature Importance Img

Type:
dataset

Dataset Type:
PlotlyDataSet

File Path:
...ta/08_reporting/plot_feature_importance_img.json

Tags:
-

Run Command:

```
kedro run --to-
outputs="plot_feature_importance_img"
```

DVC (Data Version Control) and Kedro

CONCLUSION

DVC and Kedro are two tools very similar.

But

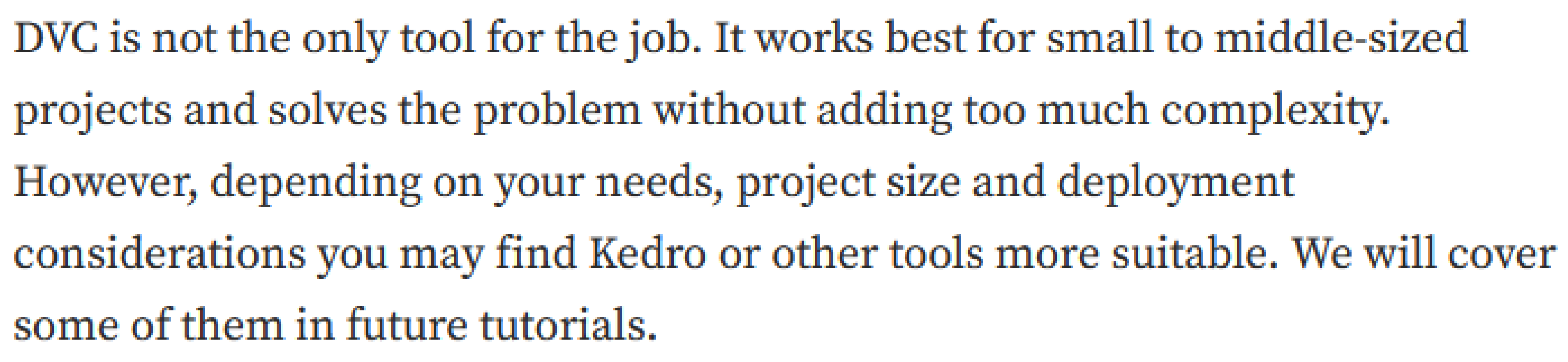
- DVC work very well with GitHub actions, because more of features are based on command-lines. This allows easy comparisons between branches of a GitHub project.
- Kedro-Viz open a browser page with all pipeline, that include node and input/output data. For all these is write the corrispective path and command-line for show or run. Kedro-Viz show in a easy way also plot, metrics and show experiments history.

REFERENCES

https://kedro.readthedocs.io/en/stable/index.html
https://dvc.org/doc/start

DVC (Data Version Control) and Kedro

From: https://medium.com/y-data-stories/creating-reproducible-data-science-workflows-with-dvc-3bf058e9797b



DVC is not the only tool for the job. It works best for small to middle-sized projects and solves the problem without adding too much complexity. However, depending on your needs, project size and deployment considerations you may find Kedro or other tools more suitable. We will cover some of them in future tutorials.