DBMS Final Project

Group3 - CapsLock快壞了

組員：林怡萱、羅敏宏、黃禹翔、廖政華

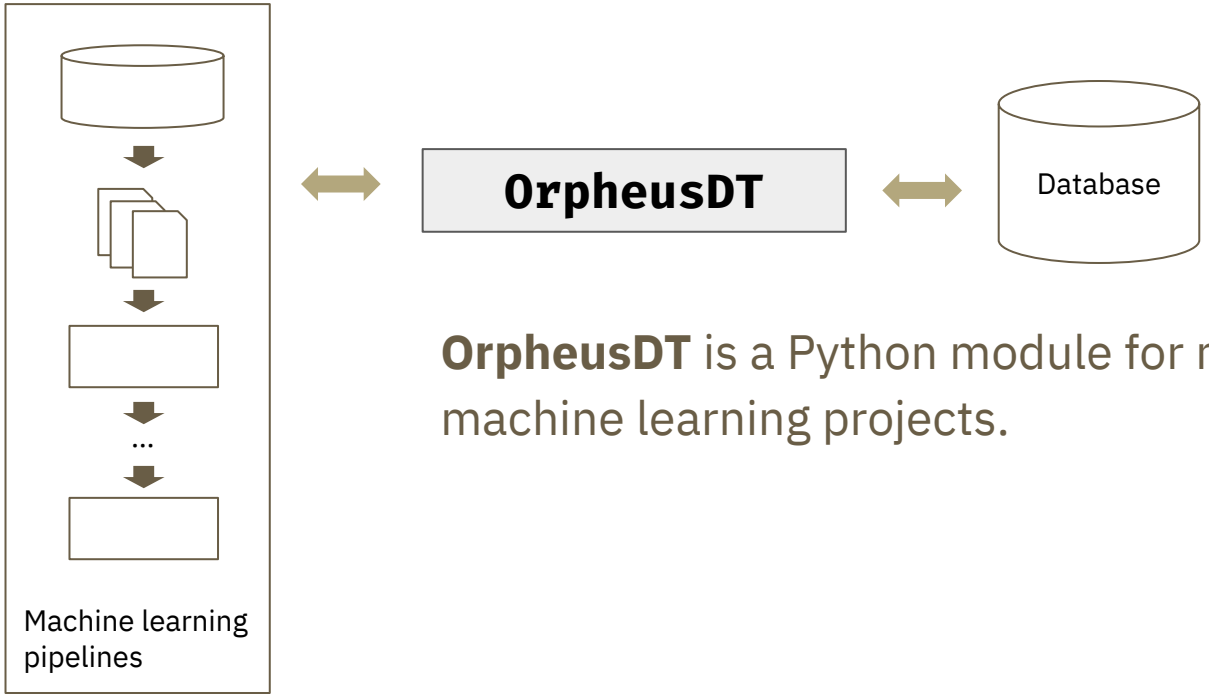
Outline

- Project overview
- Project architecture
- Project demo
- Summary

Outline

- Project overview
- Project architecture
- Project demo
- Summary

Project overview



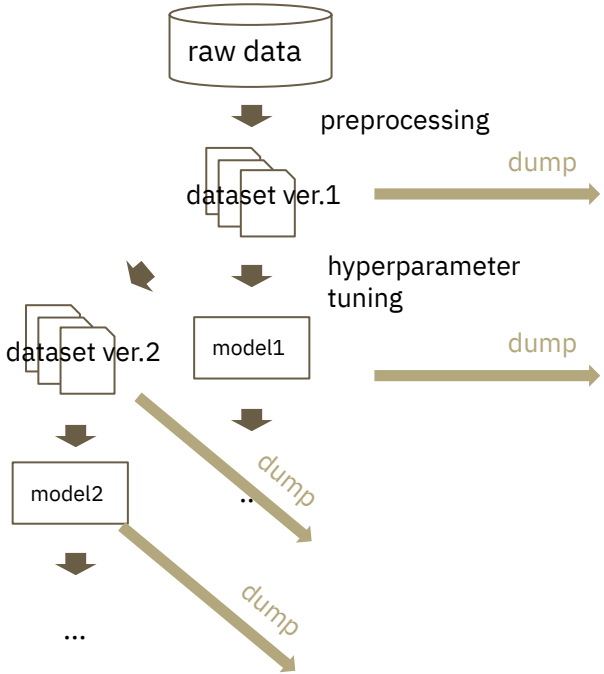
OrpheusDT is a Python module for managing machine learning projects.

Powered by

mongoDB

Motivation

The usual scenario

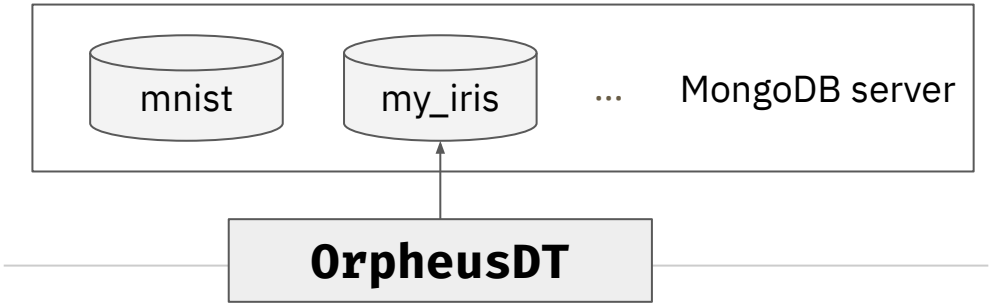


Limitations

- manual save
- limited information
- security issues with *pickle*

(WTH !?)

Our solution



.save()

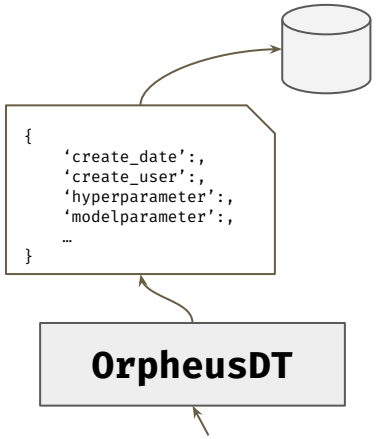
.load()
.show('list_all')
.query(precision=['glt', 0.9])
…

With OrpheusDT,
You can easily manage your projects with just a couple of simple function calls! Yay!

Our solution

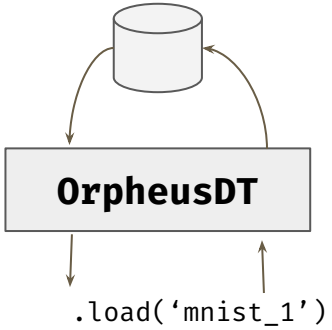
Save

OrpheusDT packs all the metadata for you.



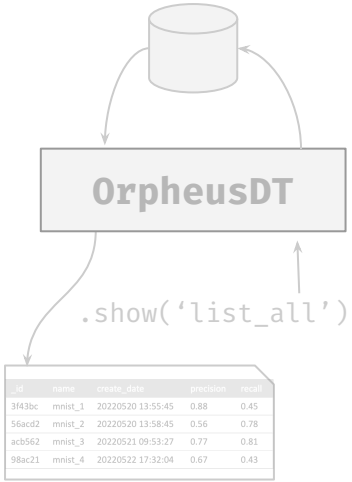
Load

Compatible with sklearn.



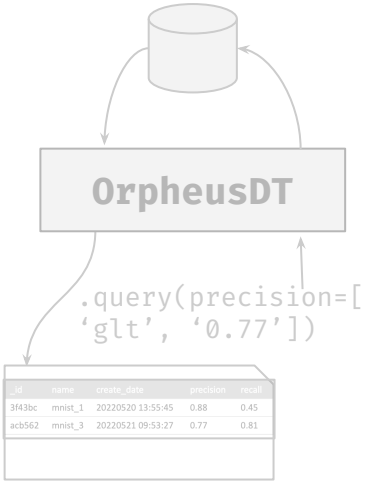
Show

Compact visualization of models.



Query

Query models with specified conditions.



Metadata

What information is recorded in metadata?

About model


```
"params": {
  "ccp_alpha": 0.0,
  "class_weight": null,
  "criterion": "gini",
  "max_depth": 5,
```

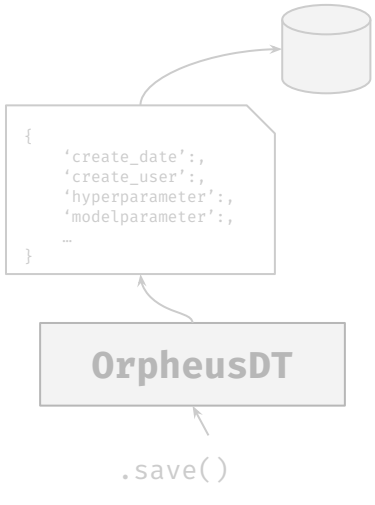

Metadata

- User
- Evaluation Scores
- Training time timestamp
- Scikit learn version
- Estimator tags (For future model selection according to the data, ex. Multiclass, allowed NaN.....)

Our solution

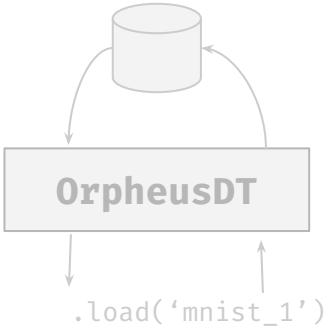
Save

OrpheusDT packs all the metadata for you.



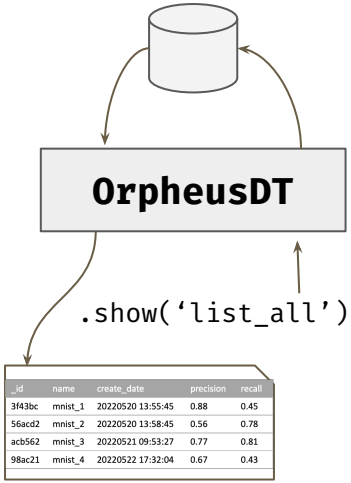
Load

Compatible with sklearn.



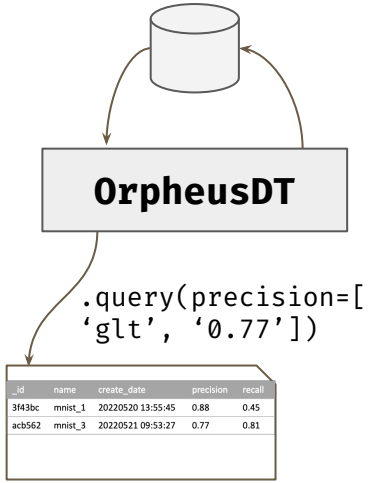
Show

Compact visualization of models.



Query

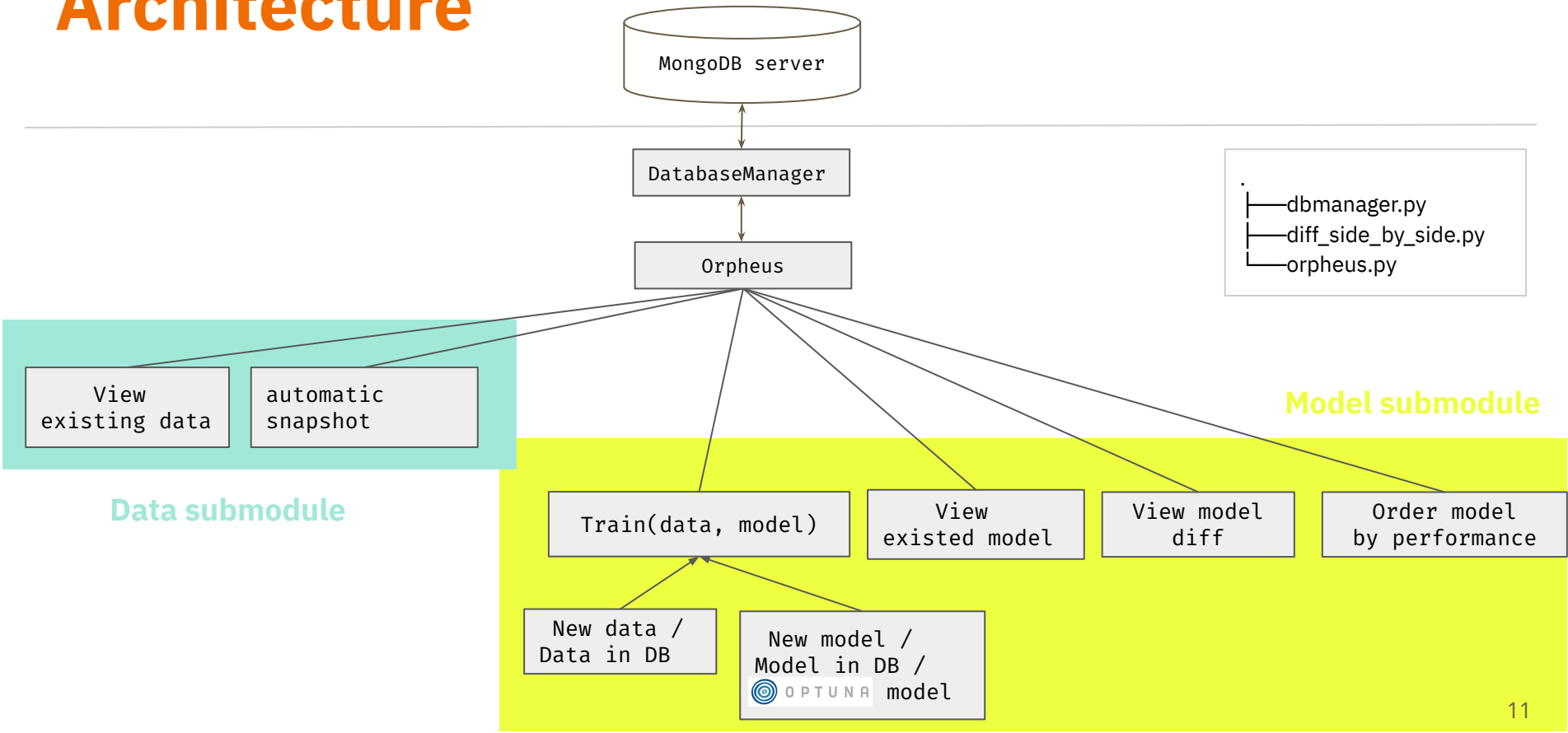
Query models with specified conditions.



Outline

- Project overview
- Project architecture
- Project demo
- Summary

Architecture



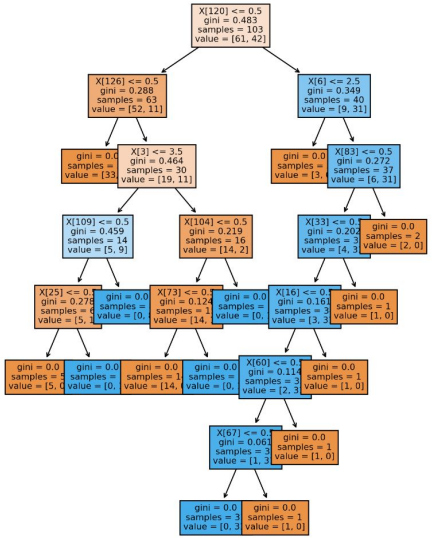
`.train()` **method**

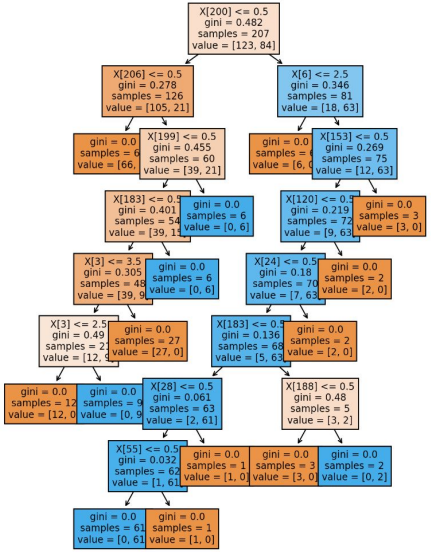
- provide function under 4 scenario through overload

new data new model	1. automatically hyperparameters tuning for fitting the model 2. save to DB
new data old model	1. continue training on old models 2. save new version model to DB
old data new model	1. train existing data to new model 2. save to DB
old data old model	review old training history and checkpoints of the model

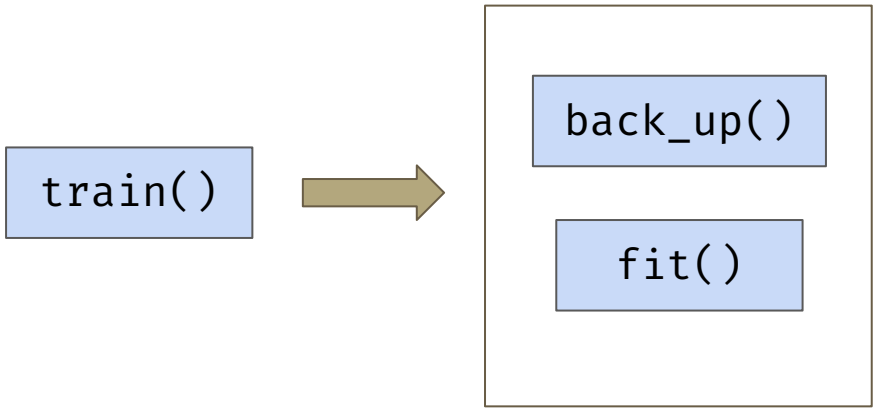
`.show_diff()` **method**

- show the difference between:
  - the lastest model in DB
  - the current version





`.back_up()` **method**



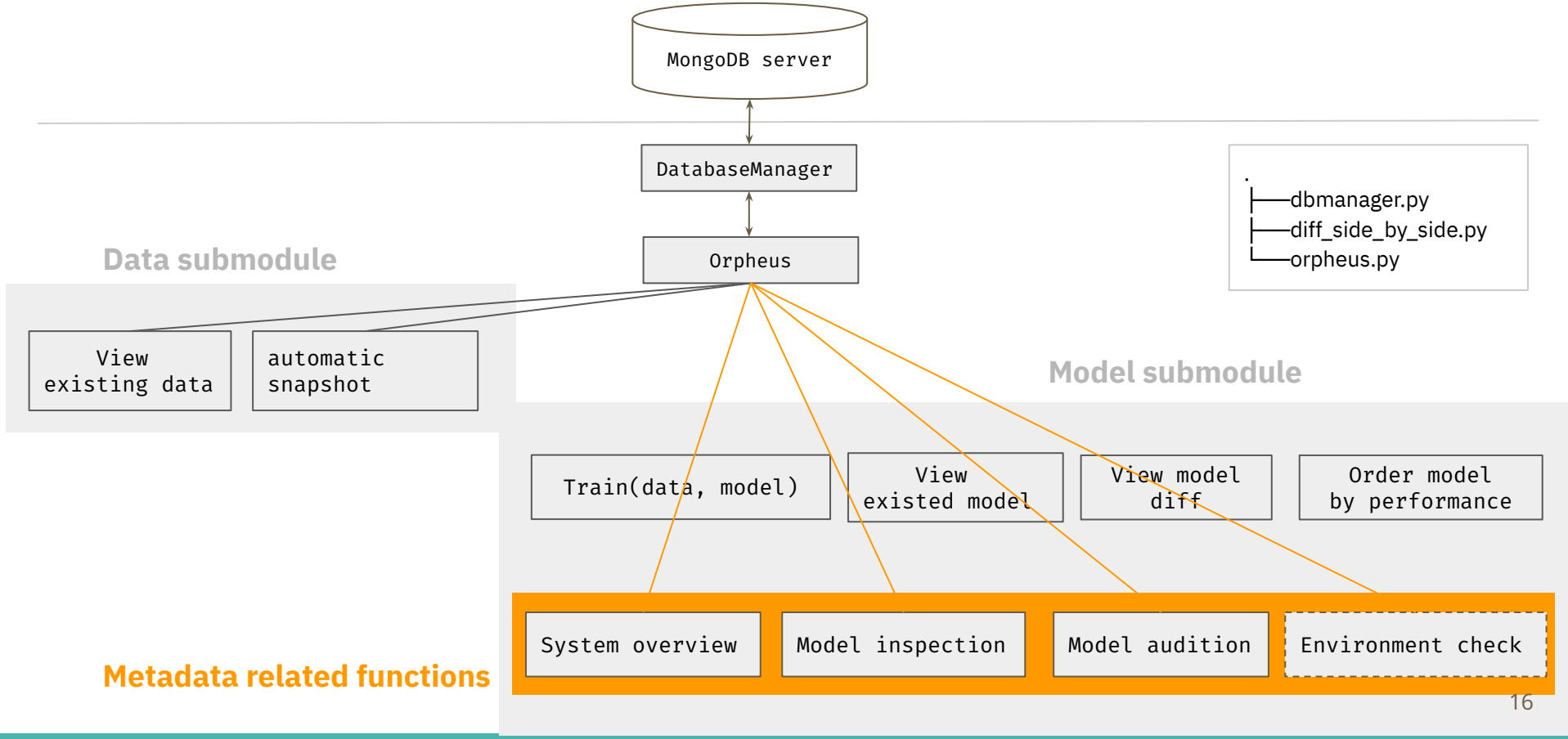
`.restore()` **method**

database crashed

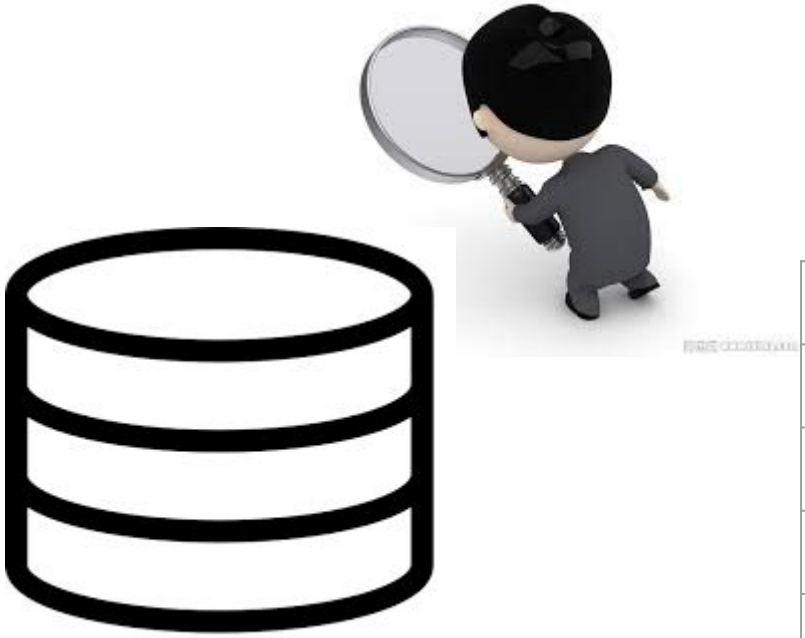
`restore()`

the latest version

Utilizing Metadata



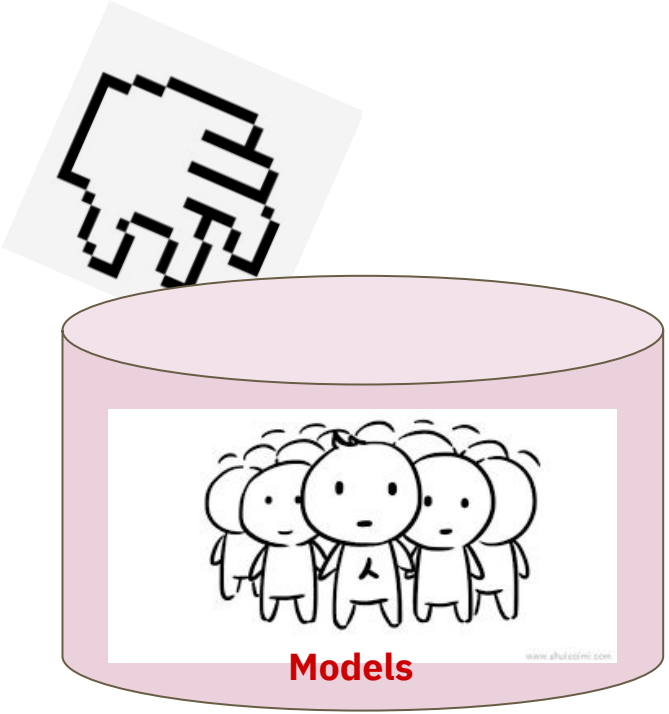
`.system_overview()` **method**



User	Timestamp	Training information	Score
Amy	2022. 5. 3	Data v1, model v2	0.83
John	2022. 4. 2	Data v2, model v2	0.75
John	2022. 3. 17	Data v2, model v4	0.92
…	…	…	…

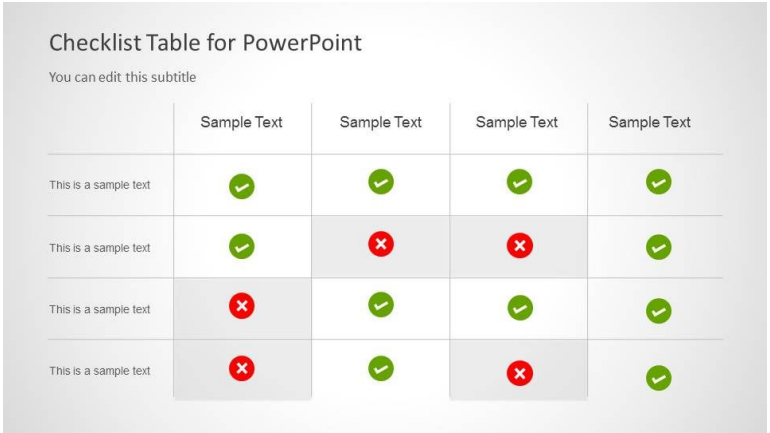
Based on **estimator tags**!

`.model_audition()` **method**

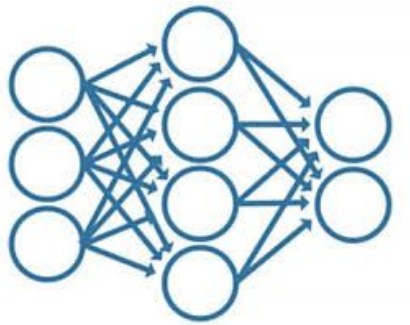


Example Requirement :

1. data contains NaN : `True`
2. Multiclass classification : `True`
3. require positive y : `False`
4. .... other `estimator tags`



Environment check (built-in mechanism)



Existed model trained under

scikit-learn version A

extract from DB

Current environment with scikit-learn version B

Remind the user that

result might be different

Outline

- Project overview
- Project architecture
- Project demo
- Summary

Outline

- Project overview
- Project architecture
- Project demo
- Summary

Summary

What we have done:

- A prototype of data management system for ML projects.
- An implementation specific to sklearn - DecisionTree model

For future:

- Integrate with more ML frameworks
- Make applicable to deep learning models
- Show more model details
  - eg. learning curve, ...

Thank you.

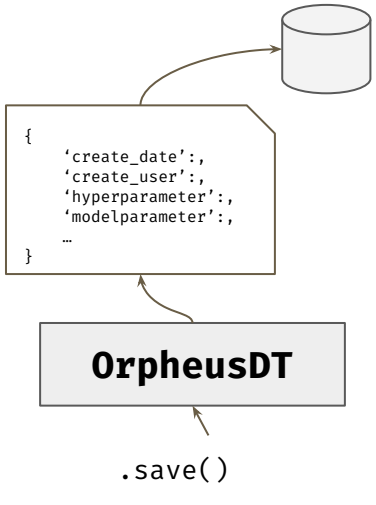
Attachment

- Github Repository
  - https://github.com/hyusterr/orpheusDT
- Oral representation video
  - https://drive.google.com/file/d/1V5n57PIxlIc6vPEpByQOwwKzR19YYXsF/view?usp=sharing

Our solution

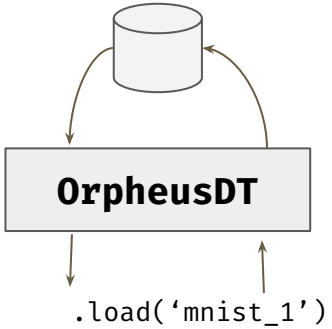
Save

OrpheusDT packs all the metadata for you.



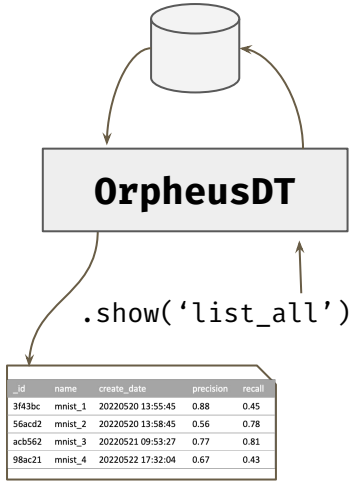
Load

Compatible with sklearn.



Show

Compact visualization of models.



Query

Query models with specified conditions.