gingado: a machine learning 🤖 library focused on economics and finance

/ʒĩ.ˈga.du/

IFC-Bank of Italy workshop on “Data science in central banking”, Part 2

What is *gingado*?

- Open-source machine learning library written in Python (under development 🛠️)
- Objective: broaden the accessibility of state-of-the-art models to a wide range of practitioners in economics and finance
- Main features:

1. Automatic benchmark models
2. Automatic data augmentation
3. Automatic documentation

… all of that with a simple API

No, really. I meant, what is “*gingado*”?

- Brazilian Portuguese
- The swing of bodies during sports, dancing, fighting
- Also related to flexibility when facing obstacles
- Analogy to the prominence of business and financial cycles
- Also, an encouragement to keep flexible & try new things

Automatic benchmark models

- Once the model and the dataset are defined, *gingado* automatically trains a benchmark model (unless asked not to train it)
- Benchmark models are useful to compare results from user attempts
- Worse case scenario, if all your attempts fail at beating the benchmark, at least you have a reasonable model

Automatic data augmentation

- “Data augmentation” means to append more data to your dataset. This is known to generally improve the performance of ML models.
- For example, in ML models working with image, data augmentation involves flipping, cutting, zooming in or out, etc.
- In economic/financial datasets, augmentation involves adding other information related to the observations in the dataset
- For example, if the dataset is a country-level panel data, data augmentation would add a lot of publicly-available data for the countries in the dataset
- New data sourced using sdmx and other APIs from official sources

Automatic data augmentation

- “Data augmentation” means to append more data to your dataset. This is known to generally improve the performance of ML models
- For example, in ML models working with image data, augmentation involves flipping, cutting, zooming in or out, etc.
- In economic/financial datasets, augmentation involves adding other information related to the observations in the dataset
- For example, if the dataset is a country-level panel data, data augmentation would add a lot of publicly-available data for the countries in the dataset
- New data sourced using sdmx and other APIs from official sources

⚠️ warning ⚠️

data augmentation in economics / finance is empirical only, as literature on data augmentation for these cases is scant.

Automatic data augmentation

Ensuring the data augmentation works

1. user defines the variables of interest for augmentation: time + geographies
2. *gingado* fetches data from BIS, IMF, ECB, Eurostat, & other sources
3. a new benchmark model is run with *all* data and it is compared to the original
4. all of the augmented data is kept if performance improves
5. if performance does not improve, only the variables that have low correlation to all of the variables in the origination dataset are kept

Automatic model documentation

- Model card inspired by Mitchell et al (2019)
- “Meta-data” about the model
- Transparency about:
  - envisage use contexts
  - how the model was evaluated, etc
- Important in model development, management, audits
- *gingado* uses JSON to store the raw model card info

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Source: Mitchell et al (2019)

Backend

- Backend based on *fast.ai* library 
- *gingado* inherits the following characteristics from *fast.ai*:
  - simple yet hackable
  - flexibility to include non-numeric datasets (texts, etc)
  - production-ready results
  - excellent community support for any *fast.ai*-specific issues
- Extensive use of 

as well

API

Designed to be as simple as possible and promote experimentation

```
import gingado

df, model_card_JSONinfo = gingado.example_data(model_card_info=True)

model = GingadoModel(
  data=df,
  ynames=['gdp_growth'],
  model_card=model_card_JSONinfo
)

model.predict(h=12) # predict GDP growth up to 12 quarters ahead
model.save("model.gingado")
```

Wishlist

“Dear Santa…”

1st priority:

- get feedback from users on the beta and release an alpha, hopefully in 2023

2nd priority:

- Include techniques for Causal ML, such as Sharma and Kiciman (2020)
- Include explainability algorithms, such as Lundberg and Lee (2017)

3rd priority:

- Integration of ML with traditional econometric techniques (eg, VAR)

If you are interested…

Please use it and share your experience!

- Open up an “issue” on GitHub if you experience any bug
- You may also open up “issues” for feature requests (please be as specific as possible)
- Contributors are welcome! Please get in touch before opening a pull request

https://github.com/dkgaraujo/gingado/

Thank you for your attention!