Marco Edward Gorelli

---

python-pandas core developer / maintainer
Samsung “Advanced Software Engineering (C++)” award holder
Bayesian Data Scientist
University of Oxford MSc
Kaggle competitions expert
PyData London Meetup assistant organiser

Education

2016-2017 **MRes, Mathematics of Planet Earth**; 2016-2017; Imperial College & University of Reading; Distinction

2015-2016 **MSc, Mathematics and Foundations of Computer Science**; 2015-2016; University of Oxford

2011-2015 **BSc, Mathematics with Professional Practice**; Brunel University London; First Class Honours Dissertation: The Tutte Polynomial and the Merino-Welsh Conjecture

Foster Award for “exceptional mathematical ability”, Level2 Award for highest grades

Open source

pandas; Aug19 - now

Data wrangling platform for Python widely adopted in the scientific computing community. I’m currently a core developer, and highlights from my contributions include:

- contributing new features (such as the `DataFrame.to_markdown` method featured in the v1.0.0 release notes);
- fixed bugs in many different parts of the core code base - several of which to do with the `Categorical` class;
- set up `pre-commit` to run linting and formatting checks during continuous integration, adding several new hooks to catch common errors;
- mentored many new contributors (more of this in the *Outreach* section)

PyMC3; Sep20 - now

Python package for probabilistic programming. My contributions include:

- migrating continuous integration pipelines from TravisCI to GitHub Actions;
- increasing test coverage;
- fixing bugs;
- enhancing/restructuring documentation;
- maintaining the `pymc-examples` gallery of Jupyter Notebooks used in the documentation;
- mentoring new contributors.

nbQA; Jul20 - now

Adapter to run any standard Python code quality tool on a Jupyter Notebook. I co-authored this Python tool and released it with the permissive MIT Licence. It's been mentioned on the Python Bytes and Talk Python to Me podcasts and is used by several open source projects (PyMC3, pyhf, pandas-profiling, sktime, and more) and was featured on the podcast Python Bytes. Source code: `https://github.com/nbQA-dev/nbQA`.

other;

I have made assorted contributions to other libraries, including:

- implementing the `types_or` method for `pre-commit`;
- fixing longstanding bug in `matplotlib` regarding the text offset not being set correctly when using scientific notation.

Experience

Data Scientist at Samsung R&D Institute UK; Nov18 - now

- Mentoring: gave workshops to follow employees on:
  - test-driven development with `pytest`
  - managing and sharing git commit hooks with `pre-commit`
  - static typing in Python with `mypy`
  - a workflow for productively using `git`
  - SHAP values for ML model interpretability
- ML sales forecasting: I was tasked with improving an ML sales forecasting model. I identified bugs in the pipeline and then, with a significantly simpler approach, achieved:
  - 5 percentage points at both monthly and yearly level, according to company's custom metric;
  - pipeline ran over 30x faster;
  - model was far more interpretable due to its simpler nature.
- Web scraping: project involved scraping data in order to make forecasts. My contributions included:
  - Implementing a Streamlit dashboard for visualising scraped data and forecasts;
  - Writing a custom Bayesian forecasting model for making predictions, allowing for inclusion of domain knowledge and quantification of uncertainty;
  - Developing a Python package to save ScrapingHub, Twitter, and third-party data to MongoDB, allowing team to reliably have access to data for insightful dashboard.
- Remote monitoring: team received data at regular intervals from remote sensors. My work included:
  - Debugging and maintaining PySpark data engineering pipeline, finding+fixing numerous inconsistencies in the process and thus improving reliability
  - Writing `pytest` test suite and continuous integration workflows from scratch
  - Writing anomaly detection module which saved data to MongoDB, thus providing project's end-users with actionable insights
- Wearable application: co-worker had written Python ML model, which another co-worker had implemented in C++. However, I was told that it did not work reliably once deployed. Hence, I:
  - Debugged the Python and C++ code, fixing numerous inconsistencies
  - Re-defined evaluation metric, aligning model accuracy with real-world accuracy (previously the evaluation metric gave overly optimistic results)
- Passed Samsung's *Advanced Software Engineering Test (C++)* (£500 reward)

Data Scientist at Sedex; Jun18 - Nov18

- Wrote custom model to work on survey data, beating third-party's benchmark
- Migrated JasperSoft dashboards to Tableau

Data Scientist at Sensium; Jan18 - May18

- Rewrote internal MATLAB visualisation tool in Python, "exceeding expecations"

Maths tutor at Oxford Exclusif Tutorial Agency; Oct18 - Jan18

Risk Analyst Intern at General Electric Capital International; Jun13 - Jun14

Outreach

- PyData Global 2020, PyData Amsterdam 2020:
  I hosted `pandas` sprints at both these events, where I mentored ~30 new contributors on how to contribute to open source and subsequently reviewed their contributions;
- Beginners Workshop on Python for Data Analytics; Jan20
  I co-mentored 40+ attendees from underrepresented minorities in tech, teaching them data analytics skills (pandas basics, pandas operations), answering questions, and providing encouragement;
- Contributing to pandas for beginners; Dec20
  Event organised by Data Umbrella, targeted at potential new contributors from underrepresented minorities in tech. I gave a talk on how to contribute to pandas and then personally assisted attendees in setting up their development environments.

---

marcogorelli@protonmail.com • `https://github.com/MarcoGorelli`