Marco Edward Gorelli

Data Scientist
University of Oxford MSc
pandas maintainer / core contributor
Kaggle competitions expert

Education

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

Courses: Computational Stochastic Processes, Partial Differential Equations (80), Dynamical Systems (84.2), Data and Uncertainty (63.1), Numerical Methods (67.13)

Dissertation: Modelling the cloud and snow surface via KPZ equation (78.6)

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

Courses: Machine Learning (64), Categories Proofs and Processes (66), Quantum Computer Science (76), Networks (70), Computational Game Theory (70)

Dissertation: Deductive verification of the s2n HMAC code (65)

2011-2015 **BSc, Mathematics with Professional Practice**; Brunel University London; First Class Honours

Dissertation: The Tutte Polynomial and the Merino-Welsh Conjecture (91)

Foster Award for "exceptional mathematical ability", Level2 Award for highest grades

Experience

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

- 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 were:
  - Wrote custom Bayesian forecasting model for making predictions, allowing for inclusion of domain knowledge and quantifying uncertainty
  - Wrote package to save ScrapingHub, Twitter, and third-party data to MongoDB, allowing team to reliably have access to data for insightful dashboard
- Remote monitoring: project involved remote sensors, from which we received data at regular intervals. My work included:
  - Debugging and maintaining PySpark data engineering pipeline, finding+fixing numerous inconsistencies in the process and thus improving reliability
  - 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 which I then wrote a test-suite for
  - Re-defined evaluation metric, aligning model accuracy with real-world accuracy (previously the evaluation metric gave overly optimistic results)
- Mentoring: gave workshops to follow employees on:
  - test-driven development
  - easier code review via pre-commit
  - Bayesian methods
  - Markov chains for denoising images
  - SHAP values for ML model interpretability
  - static typing in Python
- 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

Open source

pandas; Aug19 - now

Pandas is data wrangling platform for Python widely adopted in the scientific computing community. I've been contributing bug fixes and enhancements since August 2019, and in August 2020 was invited to join the core team. Some highlights of my activity include adding the new method `DataFrame.to_markdown` (featured in the **v1.0.0 release notes**) and hosting a pandas sprint at **PyData Festival Amsterdam**, where I mentored 15 attendees

nbQA; Jul20 - now

Open source code quality tool for Jupyter notebooks which I wrote. It's currently download more than 600 times per month, most notably by probabilistic programming framework **PyMC3** which uses it as part of its continuous integration process. Source code: `https://github.com/nbQA-dev/nbQA`.

Other projects

- Kaggle *Tweet Sentiment Extraction* competition:
  Task was to predict which parts of sentences were responsible for the given sentiment classification. I fine-tuned a pre-trained PyTorch RoBERTa NLP model, obtaining a top 8% score (bronze medal).
- Kaggle *M5 Forecasting - Uncertainty* competition:
  Task was to predict quantiles for Walmart sales. I blended together different variations of quantile regression Keras deep learning models, obtaining a top 8% score (bronze medal).