ANDREA DI FRANCIA

Data Scientist with 7+ years of expertise in public and private sector roles, utilizing statistical and machine learning models, including LLMs and Gen-AI solutions, to derive insights and build data-driven products. Skilled in Python, R, SQL, extracting insights from unstructured data, and productionising ML models leveraging Docker and Airflow.

@ andrea.l.difrancia@gmail.com | andreadifra.github.io | andreadifrancia | London, UK

EXPERIENCE

Senior Data Scientist
Health Economics and Outcomes Research (HEOR)

Oct 2024 – Present | Remote, UK

- Led meta-analysis of global meningococcal carriage prevalence using **hierarchical mixed-effects logistic regression models** with natural cubic splines to quantify prevalence by region and age group.
- Conducted **real-world evidence analysis** on a dataset of **161,000+ patients** with Type 2 Diabetes to explore BMI at diagnosis and subsequent cardiovascular and kidney outcomes.
- Implemented **age-adjusted parametric survival models** stratified by BMI subgroups to estimate 10-year cumulative incidence of health outcomes.

Senior Data Scientist
UK Health Security Agency

Jan 2022 – Sep 2024 | London, UK

- Developed and deployed **Text Classification models** using **Transformer-based LLMs** that extract hepatitis markers status from unstructured clinical notes, reaching **95%+ accuracy**. Models integrate into clinical workflows, reducing manual review burden by **10 hrs/month**.
- Implemented a **Bayesian statistical model** using ***R*** and ***Stan*** to estimate the effectiveness of the NHS Covid-19 App in **reducing positive Omicron cases** by approximately **1m**, and around **40k hospitalisations** & **3k deaths**.
- Engineered a ***Python***-based **ETL pipeline** to automate the extraction, enrichment, and ingestion of wild bird line-list data from tabular and PDF sources. reducing manual processing by **5 hrs/week** during peak avian influenza seasons, enabling efficient and timely sharing of surveillance data.
- Designed and fitted **non-linear least squares regression models** to genomic sequencing data in ***R*** in order to estimate the prevalence of different sub-lineages of the Omicron variant of SARS-CoV-2 over time.
- Partnered with University of Oxford academics to evaluate and measure the impact of the NHS Covid-19 App through **published research**.
- Leveraged cloud platforms (**AWS, Azure**) to develop & productionise analysis efficiently at scale, employing tools such as **Docker** and **Airflow**.
- Led projects and line-managed junior data scientists in cross-functional delivery team.

Economic Data Scientist
Department for Education

Apr 2017 – Dec 2021 | London, UK

- Developed ***R***-based **forecasting models** valued at **£24bn** for teacher pay-bill spending, with interactive ***shiny*** dashboard supporting policy development across the department.
- Quantified economic benefits for a successful **£240m budget proposal** for the Early Career Framework policy and created ***R-shiny*** dashboard tracking key performance metrics that received an award for outstanding achievement.
- Collaborated on **Schools' Cost Pressures Model** forecasting **£40bn** in expenditure and developed the **Apprenticeship's Levy Model**, improving accuracy in projections of **£100m** in levy spending.
- Led analysis on capital expenditure and financial sustainability for the Higher Education sector in England (**£4.3bn**), quantifying market failure risks and rationale for government intervention.

SKILLSET

Python | SQL | R | NLP | LLMs | RAG | Mixed-effects Models | Docker | PyTorch | AWS | Deep Learning | Airflow | Clustering | PowerBI | Survival Analysis | Time Series

EDUCATION

MSc. Applied Statistics & Operational Research
Birkbeck, University of London

Grade: **Distinction**

- Top 10% in cohort; highest marks: Stochastic Systems (94%), Statistical Learning (87%)
- 2-year part-time study programme; completed whilst working full-time

BSc. (Hons) Economics
University of Nottingham

Grade: **1st Class Honours**

PUBLICATIONS

Journal Articles

- L. Ferretti et al., "Digital measurement of SARS-CoV-2 transmission risk for precision epidemiology," *Nature*, 2023. [Online]. Available: https://rdcu.be/dvLyt.
- M. Kendall et al., "Epidemiological impacts of the NHS COVID-19 app in England and Wales throughout its first year," *Nature Communications*, vol. 14, 2023. [Online]. Available: https://rdcu.be/c6yas.

INTERESTS

Computer Vision
Spearheaded a hackathon project focused on automating the extraction of tables and handwritten text from a backlog of source documents containing disease notifications. Leveraged Tika and Tesseract for robust object recognition on images, facilitating data extraction for subsequent analysis and integration into downstream applications.

Mathematics
Completed half of the required modules (60 UK CATS) towards the Graduate Diploma in Mathematics offered by the University of London (academic direction from London School of Economics).