TOMAS **PEREIRA DE VASCONCELOS**

tomas@tpvasconcelos.com   tpvasconcelos.com   linkedin.com/in/tpvasconcelos   github.com/tpvasconcelos

👨🏽‍💼 **Professional profile** (summary)

- Experienced **Senior Machine Learning Engineer** and reliable **Technical Lead** with a proven track record of transforming complex Data Science problems into high-quality production-ready software solutions.
- Having worked in the software industry for **+8 years** and on a broad range of machine learning projects, I have developed a strong understanding of the data science lifecycle and the challenges that come with it.
- From machine learning, **MLOps**, software, and data engineering, to **solutions architecture** and **systems design**... I have worn many hats while helping teams solve real business problems and deliver value-driven solutions to ML problems using techniques ranging from **time-series forecasting**, **recommender systems**, **reinforcement learning**, **anomaly detection**, **Bayesian inference**, and many others...
- As a **Tech Lead** and a Senior Software Engineer, my main responsibility is ensuring my teams are able to iterate quickly *without* breaking things! I advocate and educate Data Science teams on the importance of **software engineering principles and best practices**, and how to refactor safely and efficiently.

👨🏽‍🎓 **Academic experience** (summary)

- **First Class Honours degree in Physics**, followed by a **graduate degree in Mathematical Physics**.
- **Research experience** in computational fluid dynamics and <u>a publication</u> in theoretical statistical mechanics.

Professional experience

Nike / Tech Lead Senior Machine Learning Engineer (Consultant) / Amsterdam, Netherlands / Aug 2022 - Aug 2024 · 2 yrs

My responsibility was to guide and lead all technical aspects of our forecasting system and its development, including solutions architecture, systems design, infrastructure and reliability considerations, CI/CD, speed of iteration, down to repository structure, code patterns, software tooling, and developer experience.

Product: Delivery of reliable insights on Nike's consumer demand to senior ELT members and multiple store planning and merchandise financial planning teams around the globe on time and on a regular basis.

Deliverables: Medium- to long-term forecasts for multiple metrics, various granularity levels (both temporal and hierarchical), and under different assumptions for leading demand indicators such as promotional plans.

Forecast reconciliation: Our solutions ranged from highly granular bottom-up forecasts using many contextual drivers, to high-level top-down forecasts driven by macroeconomic projections. In all cases, we had to ensure that our forecasts (as well as the forecasts produced by other teams) were aligned and consistent.

Forecast evaluation: To build trust with our business stakeholders and help our team iterate faster and more confidently on new model versions, I put forward a set of guidelines and design principles for implementing robust, unbiased, and theoretically sound evaluation strategies for our modelling approaches. First shared as an RFC within the team and later externally, I documented important considerations such as choosing the right <u>measures of forecast accuracy</u>, CV design choices, aggregation over temporal and hierarchical granularities, the difference between model-selection and evaluation and its consequences, etc.

Orchestration: Our pipelines were orchestrated and scheduled using <u>Airflow</u>, <u>Kubernetes</u>, and <u>SageMaker</u>. We defined many ETL pipelines but made all inference jobs idempotent and easy to debug by providing clear data provenance and allowing engineers to resolve issues with upstream data sources asynchronously.

Data governance: Dataset validation checks were implemented for all upstream and downstream datasets and most intermediate transformations. Apart from allowing us to better document all our datasets, this has proved useful many times in detecting upstream regressions and changes that could go unnoticed otherwise.

BEAT / Senior Machine Learning Engineer / Amsterdam, Netherlands / Jun 2021 - Jul 2022 · 1 year 2 months

My mission at BEAT was to develop, deploy, and maintain Data Science and Machine Learning solutions to detect and prevent fraud, reduce financial losses and abuses, and ensure a safe environment for the all users on our platform. All of this while keeping the business metrics healthy and the company growing!

Feature store deployment: Coordinated and led the design and implementation of a feature store solution to serve all ML and Analytics teams at BEAT. Started by testing and evaluating managed solutions such as Databricks, SageMaker, and Tecton, but settling for Feast (open-source) as the registry and serving layers. Deployed the end-to-end solution on our Kubernetes cluster using our data lake and Trino as the offline store, and a low-latency gRPC service and ElastiCache as the online store. Worked closely with Feast's dev team by providing feedback and contributing to the open-source project (see *"§ Open-source"* below).

Fraud prevention: Worked on the design and development and led the deployment of a chargeback fraud prevention system by **1)** designing the solution's architecture; **2)** validating the assumptions and performance estimations; **3)** developing an ML batch workflow (PySpark + Argo Workflows) that computes risk scores and pushes them to a Kafka topic, with asynchronous ingestion to a fast-store by a backend system; **4)** configured monitoring, alerting, and data validation for the application and infrastructure with Prometheus and Grafana, and; **5)** designing and running an online controlled experiment to measure to *real* impact of the model.

Fraud detection: Collaborated with another Data Scientist on a fraud detection problem where the labels were sparse, biased, noisy, and mostly positive and unlabelled (PU). We considered and explored several modelling options from naive binary classification with gradient boosting models to more robust active-learning approaches. The final implementation is based on an iterative semi-supervised learning solution that yields twice as many fraud cases as the previous system while keeping precision extremely high.

Tiqets / Machine Learning Engineer / Amsterdam, Netherlands / Jan 2019 - May 2021 · 2 years 5 months

As an ML Engineer, I was part of Tiqets' core Data Team. Working closely with business analysts, data engineers, product owners, and ELT members, I applied software development, data analytics, and machine learning to scale and operationalise statistical models and make the whole organisation more data-driven.

Time series forecasting: Operationalised and automated demand forecasting at Tiqets by developing a generalised time-series forecasting framework from scratch that would support pre-processing, model selection, evaluation, and periodic batch inference jobs for various business contexts and requirements. Each task was distributed across an array of Celery workers deployed on our Kubernetes cluster. Forecasted values and model metadata were pushed to Amazon Redshift and visualised by business stakeholders in our Looker BI instance. We also used DataDog for application and infrastructure monitoring.

Recommender system: Improved recommendations across our platform by integrating with the AWS Personalize service and developing robust heuristics for cold-start instances (taking distance, popularity, and seasonality into account). To help the team iterate faster and with greater confidence, we also implemented a time-dependent offline evaluation for recommender systems, curated for our e-commerce setting.

Learning-to-Rank & Reinforcement Learning: Implemented online Bayesian Reinforcement Learning-to-Rank bandit strategies (e.g. Thompson sampling), which use *explore and exploit* to continuously learn and improve the rankings of product variants on product pages. An Airflow pipeline was scheduled to frequently update all item rankings by re-sampling from updated posterior distributions as new data came in.

Supervised Learning-to-Rank: Improved product rankings by framing the task as a Supervised Machine-Learned Ranking problem, and comparing predicted rankings to a defined *ideal* ranking (*e.g.,* using nDCG).

Accelogress / Machine Learning Engineer (part-time) / Gildford, UK / Jun 2016 - Mar 2018 · 1 year 10 months

Worked closely with the CEO and Lead Developer at Accelogress on the Save-a-Space project, where I led the development and implementation of time-series forecasting models and scheduled batch inference jobs for predicting car park availability for multiple locations around the UK. I also developed and deployed a REST API to expose historical, real-time, and forecasted availability to our mobile app and web dashboards.

Open-source

I am the author of **ridgeplot**[157]★, a Python package for interactive ridgeline plots. So far, the project has had up to <u>102k PyPI downloads</u> in a month, is used by <u>over 150</u> public GitHub repositories, and is a dependency of notable projects such as <u>Shiny for Python</u>[1.3k]★, <u>Ploomber</u>[3.5k]★, and <u>NiMARE</u>[184]★. While maintaining this library, I have implemented <u>feature</u> <u>requests</u>, <u>answered</u> <u>questions</u>, and solved <u>bug</u> <u>reports</u>. The project tries to follow Python best practices, has a comprehensive <u>user guide</u> and <u>API documentation</u>, strict linting rules and type checks, high <u>test coverage</u> for both functional and unit tests, and robust CI/CD workflows.

Contributions to other projects

I try to actively contribute to open-source projects in the Python and scientific computing ecosystems, with <u>significant</u> <u>contributions</u> to high-profile projects such as <u>sktime</u>[8k]★ and <u>Feast</u>[5.6k]★ and <u>smaller</u> <u>contributions</u> to projects like <u>Apache Spark</u>[40k]★ and <u>pandera</u>[3.4k]★. My contributions span from <u>identifying</u> and <u>fixing</u> critical <u>bugs</u>, to implementing substantial improvements like <u>speeding-up checkov</u> by up to 70%. Beyond code contributions, I also <u>participate</u> in the open-source community through <u>code</u> <u>reviews</u>, <u>helping users</u> <u>troubleshoot issues</u>, and providing <u>context</u> and <u>feedback</u> where needed.

Research & Publications

<u>The second virial coefficient of bounded Mie potentials</u> / The Journal of Chemical Physics / 2017

Derivation of analytic expressions for the second virial coefficient of bounded Mie potentials.

Soft-Core Lennard-Jones model / BSc Physics Thesis / Grade: 82% / 2017

My research was focused on the mathematical properties of special interactive potential models.
Official feedback from the thesis evaluation panel:

- *"Tomas observed and brought out [...] a number of very interesting results, of a caliber which would reflect well on a researcher of many more years experience."*
- *"The series expansion and MD modeling aspects of the project were definite steps forward in our understanding of the behavior of this system."*
- *"[Tomas] derived single-handedly a series expansion in terms of orthogonal polynomials, which is a new approach as far as I am aware."*

Forecasting UK carpark availability / Poster Presentation / South East Physics Network / 2016

Education

GradD Mathematics / King's College London, UK / Sep 2017 - Aug 2018

- Developed a robust mathematical foundation for further studies and research in theoretical physics.

BSc Physics / Royal Holloway, Uni. of London, UK / Sep 2013 - Jun 2017 / First Class Honours

- Developed the intuition to identify relevant laws and principles and apply the appropriate mathematical methods and approximations when working on complex problems.
- Worked on scientific computing projects with a focus on data analysis and computational fluid dynamics.

High School (Exchange Year) / River Ridge High School, USA / Aug 2011 - May 2012 / 93% GPA

- Top scores for mathematics, physics, and chemistry with final grades ranging between 95% and 100%.

High School (Science Track) / Salesianos de Manique, Portugal / Sep 2008 - Jun 2011 / 80% GPA

- Ranked in the national top 3% for the Physics and Chemistry national exam.

Other skills / Some ATS/SEO keywords not mentioned above… 🤖

Python | SQL | Docker | Spark | DevOps | Database | Agile | Microservices | OOP | git | A/B Testing | CI/CD | Data drift | AI | Cloud