Dr. Georgios **Douzas**

Software Engineer · Machine Learning Engineer · AI Researcher

georgedouzas.github.io | georgedouzas | Georgios Douzas | Georgios Douzas

Summary

Machine learning engineer with expertise in developing ML models, building scalable systems, and automating workflows. Managed teams and projects from start to finish, ensuring timely delivery. Proficient in Python, R, C++, and cloud platforms (AWS, GCP, Azure). Skilled in TensorFlow, PyTorch, and scikit-learn. Maintainer of open-source projects like imbalanced-learn-extra and sports-betting. Published research on class imbalance, Reinforcement Learning and Generative Deep Learning. Mentored junior engineers and led cross-functional teams to achieve business goals. Recognized in Stanford's top 2% scientists globally in 2024.

Work Experience

ForthTech V.C

Athens, Greece

Chief Technology Officer and Key Executive — *August 2020 - Present*

As a stakeholder, key executive and member of Board of Directors at ForthTech, a State-funded Venture Capital Fund focused on advancing Greece's tech sector, I played a leadership role in managing a €25 million investment portfolio. Collaborating closely with my partners, I led the evaluation of high-potential startups, overseeing the technical due diligence process to ensure strategic investment decisions. I managed the creation of in-depth technical reports, assessing the feasibility, scalability, and long-term potential of startups, leveraging my expertise in artificial intelligence and software engineering. Beyond evaluation, I worked directly with entrepreneurs, providing strategic guidance to refine their technology, strengthen their business models, and position their products for growth. I set key milestones, identified technical improvements, and facilitated connections with resources and networks to accelerate their success. Additionally, I contributed to the broader development of Greece's tech ecosystem by fostering innovation and entrepreneurship, helping startups build disruptive technologies with global potential.

Trasys

Athens, Greece

Machine Learning Architect — *August 2020 - Present*

I designed and implemented the Parallel Distribution web application for the European Medicines Agency to automate PDF content comparison and reporting. I handled everything, including gathering requirements, designing the system, coding, dockerizing, and deploying it. The system uses Python, machine learning, computer vision, and Azure Form Recognizer service and moved from a proof of concept to a fully operational production tool. I also developed the PFAS web application for the European Chemicals Agency as a proof of concept to classify user and organizational comments. This project used machine learning models, RAG processes, and LLM prompt tuning and was deployed on Vertex AI Studio. Additionally, I worked on building and testing ETL pipelines to help migrate the European Chemicals Agency's main application (SDAP) to the cloud. As part of a team of 10 engineers, I focused on making sure the data integration worked smoothly and that the system could scale.

NOVA IMS University of Lisbon

Lisbon, Portugal

Machine Learning Researcher — *September 2018 - Present*

As the first author, I have published research on methods to handle class imbalance, including clustering-based oversampling techniques, Geometric SMOTE, and its extension to regression tasks. I also developed G-SOMO, which combines self-organizing maps with Geometric SMOTE to create synthetic data, improving classification performance. Additionally, I explored the use of Genetic Programming to improve data collection in offline reinforcement learning. I have also worked with deep learning models, using Conditional Generative Adversarial Networks (CGANs) for oversampling and training models with TensorFlow, Keras, and PyTorch. My research has been published in leading machine learning journals, and I have shared my implementations as open-source tools to support collaboration and accessibility. In 2024, I was ranked among the top 2% of scientists globally by Stanford University for my career contributions and publications in machine learning research.

NOVA IMS University of Lisbon

Lisbon, Portugal

Machine Learning Engineering Manager — *September 2019 - September 2021*

I designed and led the development of MapIntel, an open-source system for extracting insights from large text datasets. As the project lead, I managed a team of four developers, overseeing the entire development lifecycle from architecture design to deployment. I set technical priorities, assigned tasks, and ensured efficient collaboration to meet project milestones. MapIntel supports natural language queries, question-answering, and visual exploration using retrieval-augmented generation (RAG) engines and topic modeling. To guarantee scalability and reliability, I led the deployment strategy on AWS, utilizing services like SageMaker and Bedrock. My role extended beyond technical implementation to strategic decision-making, ensuring the platform met user needs while maintaining high performance and operational efficiency.

Tripsta

Athens, Greece

Head of Algorithmic Pricing — *October 2017 - September 2018*

At Tripsta in Athens, Greece, I led a team of four—including a data engineer, a database developer, and two software engineers—in the design, implementation, and deployment of an automated algorithmic pricing system. As the project lead, I managed the development lifecycle, setting technical priorities, and ensuring seamless collaboration across teams. The system incorporated machine learning estimators for add-ons, competitor pricing, and dynamic pricing models, as well as metaheuristic algorithms for multi-objective budget optimization. I oversaw the optimization of the system to process terabytes of training data while handling up to 50,000 requests per second, maintaining prediction times under 100 milliseconds. To achieve this, I designed the system to use technologies such as Python, Java, Scala, Spark, Dask, scikit-learn, and jMetal. My leadership contributed to improving pricing accuracy, system speed, and operational efficiency, directly enhancing the company's market competitiveness. The automated pricing system became a key component of Tripsta's pricing strategy, enabling real-time optimization and data-driven decision-making across the business.

Quantum Retail Technology

Remote

SENIOR DATA SCIENTIST

October 2016 - September 2017

At Quantum Retail Technology, I worked as a Senior Data Scientist, specializing in demand forecasting for retail inventory optimization. My role involved developing and applying machine learning techniques to analyze consumer demand, optimize stock allocation, and enhance promotional effectiveness. I was responsible for building predictive models that identified missed sales opportunities, refined inventory distribution strategies, and improved decision-making across various retail functions. I contributed directly to Q Analytics, a suite of cloud-based software services designed to provide data-driven inventory management solutions. My work focused on developing models to estimate demand fluctuations, helping retailers maximize sales potential while minimizing overstock and markdowns. I also leveraged ML-driven insights to align stock levels with real-time selling patterns, optimize product packaging, and refine network flow strategies to reduce supply chain inefficiencies. Beyond technical development, I collaborated closely with stakeholders and business leaders, ensuring that analytical insights translated into actionable strategies with measurable business impact. My work was fully integrated into the Q platform, enabling retailers to make smarter, data-driven decisions that improved profitability, efficiency, and overall operational effectiveness.

CERN

Remote

SENIOR MACHINE LEARNING ENGINEER

May 2016 - September 2016

At CERN, I worked as a Machine Learning Engineer, where I developed parallelized features for TMVA (Toolkit for Multivariate Data Analysis) within ROOT, CERN's software framework used for data analysis in high-energy physics. My work focused on enhancing big data processing, statistical analysis, and visualization capabilities for particle physics experiments. These enhancements were crucial in improving the efficiency of data handling and enabling more accurate analysis of complex datasets generated by experiments such as those at the Large Hadron Collider (LHC). One of my key contributions was implementing brute-force and metaheuristic algorithms for hyperparameter grid search in machine learning models. This process, crucial for optimizing machine learning algorithms, was initially based on a legacy C++ system. By modernizing this system and utilizing Python and Spark, I was able to significantly improve the speed and scalability of the process. The new setup allowed for the handling of vast amounts of data across distributed systems, providing faster model training and optimization. I also collaborated with the physics and engineering teams at CERN to ensure that these updates were effectively integrated into existing workflows. The improvements I made directly contributed to the analysis of high-dimensional particle physics data, helping researchers derive valuable insights from experimental results.

IRI

Greece

SCIENTIFIC SOFTWARE ENGINEER CONSULTANT

October 2014 - May 2016

At IRI Greece, I worked as a scientific software engineer consultant within the Solutions and Innovation Team (R&D), playing a key role in migrating the “Price & Promo Analytics” solution from Base SAS to Hadoop and Spark. This migration, which I helped drive from proof of concept to full production, supported over $25 million in annual revenue. I provided strategic technical guidance on parallelization methods and data processing optimization to ensure the system could efficiently handle large-scale datasets. My expertise in R was instrumental in developing scalable data pipelines, utilizing libraries like dplyr for data manipulation and foreach, doParallel for parallelized computation. I also built and validated generalized linear models to analyze competitor pricing and inform pricing strategies. Beyond migration efforts, I facilitated the transition of statistical models from SAS to Python and integrated R and Julia for specialized analytical tasks. Additionally, I developed a Shiny dashboard, enabling business users to easily interact with and extract insights from complex data. My contributions significantly improved system scalability, reduced processing times, and enhanced the overall performance of the analytics platform, making it more efficient and accessible for both internal teams and external clients.

Sports Performance Training

Greece

FOUNDER & CTO

September 2009 - September 2013

I co-founded and served as the CTO of Sports Performance Training (SPT), where I provided consulting services to professional athletes and sports organizations across various sports. I developed custom solutions using biometric sensors, signal processing, descriptive statistics, and predictive modeling to optimize training loads and improve performance during competition periods. By analyzing performance data, I delivered individualized training guidelines to clients, with a focus on maximizing their performance and minimizing the risk of sports-specific injuries.

Education

Department of Physics, National Technical University of Athens

Athens, Greece

PHD IN THEORETICAL PARTICLE PHYSICS

September 2003 - September 2008

- Democritus Institute Graduate Program Fellowship, awarded yearly after national exams

Department of Physics, National Technical University of Athens

Athens, Greece

MSC IN PHYSICS AND TECHNOLOGICAL APPLICATIONS

September 2001 - September 2003

- Ranked 1st in Class

Department of Physics, National Technical University of Athens

Athens, Greece

BSC IN PHYSICS

September 1997 - September 2001

- Ranked 2nd in Class

Publications

Intraday trading via Deep Reinforcement Learning and Technical Indicators

arXiv

DOUZAS, G. AND BACAO, F.

2024

Using Genetic Programming to Improve Data Collection for Offline Reinforcement Learning — *SSRN*
Halder, D., Bacao, F. and Douzas, G. — *2024*

Improving the quality of predictive models in small data GSDOT: A new algorithm for generating synthetic data — *PLOS ONE*
Douzas, G., Lechleitner and Bacao, F. — *2022*

Geometric SMOTE for regression — *Expert Systems with Applications*
Camacho, L., Douzas, G. and Bacao, F. — *2022*

G-SOMO: An oversampling approach based on self-organized maps and geometric SMOTE — *Expert Systems with Applications*
Douzas, G., Rauch, R. and Bacao, F. — *2021*

Increasing the Effectiveness of Active Learning: Introducing Artificial Data Generation in Active Learning for Land Use/Land Cover Classification — *Remote Sensing*
Fonseca, J., Douzas, G. and Bacao, F. — *2021*

Improving Imbalanced Land Cover Classification with K-Means SMOTE: Detecting and Oversampling Distinctive Minority Spectral Signatures — *Information*
Fonseca, J., Douzas, G. and Bacao, F. — *2021*

Imbalanced Learning in Land Cover Classification: Improving Minority Classes' Prediction Accuracy Using the Geometric SMOTE Algorithm — *Remote Sensing*
Douzas, G., Bacao, F., Fonseca, J. and Khudinyan, M. — *2019*

Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE — *Information Sciences*
Douzas, G. and Bacao, F. — *2019*

Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE — *Information Sciences*
Douzas, G., Bacao, F. and Last F. — *2018*

Effective data generation for imbalanced learning using conditional generative adversarial networks — *Expert Systems with Applications*
Douzas, G. and Bacao, F. — *2018*

Oversampling for Imbalanced Learning Based on K-Means and SMOTE — *arXiv*
Last, F., Douzas, G. and Bacao, F. — *2017*

Geometric SMOTE: Effective oversampling for imbalanced learning through a geometric extension of SMOTE — *arXiv*
Douzas, G. and Bacao, F. — *2017*

Self-Organizing Map Oversampling (SOMO) for imbalanced data set learning — *Expert Systems with Applications*
Douzas, G. and Bacao, F. — *2017*

Self-Organizing Map Oversampling (SOMO) for imbalanced data set learning — *Expert Systems with Applications*
Douzas, G. and Bacao, F. — *2017*

Coset space dimensional reduction and Wilson flux breaking of ten-dimensional N=1, E8 gauge theory — *The European Physical Journal*
Douzas, G., Grammatikopoulos, T. and Zoupanos, G. — *2009*

Coset space dimensional reduction and classification of semi-realistic particle physics models *Progress of Physics*

DOUZAS, G., GRAMMATIKOPOULOS, T., MADORE, J. AND ZOUPANOS, G. *2008*

Honors & Awards

RESEARCH

2024 **Top 2% Scientist**, Standford University global ranking

EDUCATION

2002 **2nd Place**, Democritus Institute Graduate Program Fellowship, awarded yearly after national exams

Skills

PROGRAMMING LANGUAGES

Skilled in Python, C++, Java, Scala, R, and JavaScript, with experience building reliable software for various applications. Familiar with basic programming in C, Go, and Rust, making it easy to pick up and work with new technologies.

WEB DEVELOPMENT

Experienced in building web applications using HTML, CSS, JavaScript, React, Flask, and Django. Comfortable with both front-end and back-end development, focusing on clean, maintainable code, version control with Git, and automated testing.

CLOUD SERVICES

Hands-on experience with AWS, GCP, and Azure, using tools like SageMaker, Vertex AI, Bedrock, and Azure AI Studio for machine learning projects. Skilled in deploying and managing applications on the cloud, ensuring they are scalable and efficient.

DEVOPS

Proficient in using Docker and Kubernetes to containerize and manage applications. Experienced in setting up CI/CD pipelines for automated deployment and using Terraform for managing cloud infrastructure. Familiar with tools like AWS CloudWatch, GCP Stackdriver, and Azure Monitor for logging and system monitoring.

MLOPS

Experienced in automating machine learning workflows with tools like MLflow, SageMaker Pipelines, and Vertex AI Pipelines. Skilled in deploying and monitoring machine learning models in production, ensuring reliability and performance.

DATA ENGINEERING

Skilled in building ETL pipelines with Apache Spark, Hadoop, and cloud services like AWS Glue and GCP Dataflow. Experienced in workflow orchestration tools like Airflow and Prefect to manage and schedule data pipelines. Proficient in handling both structured and unstructured data for analytics and machine learning workflows.

Open-Source

As the creator and maintainer of several open-source projects, I have contributed significantly to the Python community by developing tools that enhance machine learning, system simulation, and software development practices. Here's a brief overview of some of my ongoing projects:

IMBALANCED-LEARN-EXTRA

imbalanced-learn-extra is an extension of the popular imbalanced-learn library, introducing novel oversampling algorithms to address class imbalance in datasets. This package includes a general interface for clustering-based oversampling algorithms and the Geometric SMOTE algorithm, which handles both numerical and categorical features. The project has garnered attention for its innovative approach to oversampling algorithms, as evidenced by its growing number of stars on GitHub.

SPORTS-BETTING

sports-betting is a comprehensive collection of AI tools designed for creating, testing, and deploying sports betting models. It offers a Python API, command-line interface, and a user-friendly GUI built with Reflex, facilitating seamless integration into various workflows. The repository has been recognized for its utility in the sports analytics community, reflected in its increasing star count.

SCIKIT-COMPLEXITY

scikit-complexity is a Python package aimed at simulating and modeling complex systems. While the project is still in its early stages, it is intended to contribute to the field of complex systems by providing tools to model system behavior and analyze its complexity. The project is still being developed but has great potential for future use in both research and practical applications.

BRAINBLOCKS

brainblocks is a reinforcement learning library implementing various classical, deep, and LLM-based reinforcement learning algorithms. Although it's in the early stages of development, the project aims to become a valuable resource for exploring reinforcement learning methodologies and algorithms.

COPIER-PDM-NOX

copier-pdm-nox is a Copier template designed for Python projects using PDM and managed by Nox for multi-environment task execution. It's still in the initial development phase, but the goal is to provide a solid starting point for Python developers looking for streamlined project setup and maintenance. It focuses on best practices and simplifies managing Python projects across environments.