Cainã Max Couto da Silva

Senior Data Scientist | Data Science Specialist

São Paulo, Brazil

📱+55 11 98671-1630 | ✉ cmcouto.silva@gmail.com | 🏠 cmcouto-silva.github.io | cmcouto-silva | cmcouto-silva | cmcouto-silva

Summary

As a highly skilled data scientist with a PhD in bioinformatics and over ten years of working on relevant projects, I developed a strong data science and analytics foundation. I have spent the last few years working at world-renowned companies, developing end-to-end machine learning applications. Additionally, driven by my passion for knowledge, I've taught specialized courses in various data science topics. I am eager to apply my expertise and create meaningful impacts.

Industry Experience

Schlumberger
Houston - TX, USA (remote)

Data Scientist | World's largest offshore drilling company — *Jan 2023 - Current*

- Code refactoring and predictive modeling for machine failures, saving huge amounts of U$.
- Applied technologies: Dataiku, GCP, Python, SQL, Docker, machine learning and data visualization libraries.
- Idea for innovation using AI ranked the **top 12** out of 237 teams worldwide. I was responsible for the pitch to the CEOs.

Escola DNC
Sao Paulo, Brazil

Data Science Consultor (Freelancer) — *Oct 2021 - Current*

- Provide group mentorship and Q&A sections. Prepare test exercises and assignments from core Python to model deployment.
- Covered material: Python, PySpark, regression, classification, clustering, recommender systems, model evaluation, and model deployment.

Ambev Tech
Sao Paulo, Brazil

Data Scientist | World's largest beer brewer company — *Mai 2022 - Dec 2022*

- Build the data pipeline for automating pricing and promotion policies.
- Advanced forecasting modeling for multiple products with hierarchical reconciliation.
- Applied technologies: Databricks, Python, Pyspark, SQL, MLFlow, Scikit-learn, and specific forecasting and data visualization libraries.

Remessa Online
Sao Paulo, Brazil

Junior Data Scientist | Fintech — *Oct 2021 - Apr 2022*

- Exploratory data analysis, time series forecasting, modeling customer retention.
- Meetings with commercial and marketing teams to present insights obtained from statistical, graphical, and machine learning analysis.
- Applied technologies: Databricks, Python, Pyspark, SQL, MLFlow, Scikit-learn, Pycaret, Matplotlib/Seaborn, Plotly.

Eli Lilly and Company
Indianapolis - IN, USA (remote)

Safety Data Sciences Associate — *Jun 2021 - Oct 2021*

- Work on queries and reports for teams worldwide.

Academic Experience

M.B.A. in Data Sciences & Analytics
Mai 2021 - Aug 2023

ESALQ - Universidade de São Paulo — *São Paulo, Brazil*

- In-depth study of machine learning models.
- Developed an end-to-end hybrid ML model for churn prediction (available here)

Ph.D. in Genetics and Evolutionary Biology
Jul 2016 - Apr 2021

Universidade de São Paulo — *São Paulo, Brazil*

- Analysis, visualization, and reporting of genomic data using R, Python, and bash scripting.
- Non-supervised algorithms (*e.g.* PCA), descriptive and inferential statistics, Bioconductor R packages.
- I provided all the code and instructions for replicating my thesis in this repository (Brazilian Portuguese).

M.Sc. in Biological Sciences
Apr 2014 - Mar 2016

Universidade de São Paulo — *São Paulo, Brazil*

- I studied the role of the interaction between the proteins PrPC and STIP1 in adult neurogenesis.
- Main techniques: primary cell culture, immunofluorescence, and hypothesis testing.

B.A. in Biological Sciences
Feb 2011 - Dec 2013

Universidade Guarulhos — *São Paulo, Brazil*

- Best academic performance's Award

Publications

- **Couto-Silva, C. M.**, Shetty, S., Olid-Gonzalez, A., Wallez, G., Chatelet, C., Kohar, A. (2024). Mitigating Nonproductive Time: A Novel Algorithm for Dsl Fault Detection. OnePetro. https://doi.org/10.2523/IPTC-24515-MS .
- **Couto-Silva, C. M.**, Nunes, K., Venturini, G., Araújo Castro e Silva, M., Pereira, L. V., Comas, D., Pereira, A., Hünemeier, T. (2023). Indigenous people from Amazon show genetic signatures of pathogen-driven selection. Science Advances, 9(10). https://doi.org/10.1126/sciadv.abo0234.
- Castro e Silva, M. A., Ferraz, T., **Couto-Silva, C. M.**, Lemes, R. B., Nunes, K., Comas, D., Hünemeier, T. (2021). Population Histories and genomic diversity of South American natives. Molecular Biology and Evolution, 39(1). https://doi.org/10.1093/molbev/msab339.
- Jacovas, V. C., **Couto-Silva, C. M.**, Nunes, K., Lemes, R. B., de Oliveira, M. Z., Salzano, F. M., Bortolini, M. C., Hünemeier, T. (2018). Selection scan reveals three new loci related to high altitude adaptation in native Andeans. Scientific Reports, 8(1). https://doi.org/10.1038/s41598-018-31100-6.

Highlighted courses

Machine learning in Python with scikit-learn

France (remote)

INRIA

Mai 2022

Manipulation of NGS Data for Genomic and Population Genetics Analyses - 2nd edition

Barcelona, Spain

TRANSMITTING SCIENCE

Feb 2020

In addition, I have taken dozens of data science-related courses, mainly from DataCamp, Coursera, and Kaggle.

Technical Skills

Using the following tools, I can perform data cleaning, wrangling, and visualization, build supervised and unsupervised models, and evaluate and optimize **machine learning** models. I care a lot about storytelling and reproducible research.

PROGRAMMING

- **Python**
- **R**
- **SQL**
- **Bash scripting**

FURTHER TOOLS

- **Machine Learning libraries**
- **PySpark**
- **MLFlow**
- **GCP**
- **Git/GitHub**
- **Virtual environments**
- **Docker**

Talks

PyData Global 2023

Global (remote)

INTRODUCTION TO MACHINE LEARNING PIPELINES: HOW TO PREVENT DATA LEAKAGE AND BUILD EFFICIENT WORKFLOWS

Dec 2023

- A 2-hours workshop explaining how pipelines help to prevent data leakage and ensure model stability by allowing for proper separation of training, validation, and test data. Through a blend of theory and practice, it walked the audience through code chunks in Python using well-known open-source packages to ensure a complete understanding of the machine learning pipelines.
- Slides and code available here.

PyData SP

São Paulo - Brazil (remote)

ASSOCIATION ANALYSIS: HOW TO EXTRACT VALUE FROM CATEGORICAL DATA (TRANSLATED FROM PORTUGUESE)

Mai 2022

- Workshop highlighting data science techniques for analyzing categorical data, such as chi-square, Cramér's V, CA, MCA, entropy, information gain, and so on.
- Code available here

Additional Information

Languages

- Portuguese (native)
- English (professional)
- Spanish (intermediate)

Conferences

Throughout my academic career, I had the opportunity to participate in various training courses and international conferences, including a 3-month internship in Barcelona - Spain.