Blog: vitaliset.github.io
LinkedIn: carlo-lemos
GitHub: github.com/vitaliset

Carlo Domenico Longo de Lemos

Work Experience

Experian (credit bureau) - Latin America Datalab — 11 mo • São Paulo, Brazil

- **Senior Data Scientist** — 05/2022 - Now

– Led the development of a model that estimates the credit default sensibility to increase in credit card limit using covariate adjustment (or S-learner, using causalml meta-learner jargon).

– Currently working on insurance risk estimation.

- **Data Scientist II** — 11/2021 - 04/2021

– Leader of model development within the AgriScore team, responsible for building and maintaining a credit score aimed at the agricultural market. Activities include restructuring the out-of-time cross-validation framework, designing and testing new metrics and targets appropriate to our B2B customers, and feature engineering.

– Led two junior data scientists.

Itaú Unibanco (bank) - Advanced Analytics Products Management — 2 yr 1 mo • São Paulo, Brazil

- **Data Scientist I** — 11/2020 - 10/2021

– Consultancy to Itaú teams with projects aimed at solving business problems applying Python's Machine Learning toolkit (mostly for Causal Inference problems) and HiveQL for querying data with Hadoop.

– Organization and presentation of technical meetings on Data Science (Journal Club and MeetUps);

– Teaching Machine Learning classes in internal study groups and external training;

– Recruitment and technical follow-up (tutoring) of new team members.

- **Data Science Intern** — 10/2019 - 10/2020

Education

Master's degree in Probability and Statistics — São Paulo, Brazil
University of São Paulo (USP) - GPA: 4/4 — 08/2021 - Now

Bachelor's degree in Pure Mathematics — Santo André, Brazil
Federal University of ABC (UFABC) - GPA: 3.88/4 — 05/2015 - Now

Bachelor's degree in Science and Technology — Santo André, Brazil
Federal University of ABC (UFABC) - GPA: 3.88/4 — 05/2015 - 04/2021

Tools

- **Python:** Good understanding of object-oriented programming, applying design patterns and SOLID principles when needed. Classic data mining and visualization libraries: matplotlib, pandas, sklearn, statsmodels, tensorflow, nltk, hyperopt, skopt, imblearn, category_encoders, lgbm, xgboost, time-robust-forest, causalml, stable_baselines3, river, aif360, abc, collections, itertools, functools, toolz etc.
- **Database querying:** SQL, HiveQL, SASpy, Hadoop.
- **Others:** Git, Scala, LaTeX, p5js, Processing.
- **Languages:** portuguese (native) and english (advanced).

Interests

causal inference, machine learning fairness, out-of-distribution generalization (including dataset shift, but not only), model interpretability, statistical learning theory, online learning, (contextual) multi-armed bandits, reinforcement learning, black-box optimization, adversarial training, software engineering best practices, open-source etc.

Additional Courses

University of São Paulo — São Paulo, Brazil
Classes: *Statistical Inference* and *Bayesian Analysis* - Grade: 9.0/10 — 07/2020 - 06/2021

National Institute for Pure and Applied Mathematics (IMPA) — Rio de Janeiro, Brazil
Summer scholarship. Class: *Introduction to the Theory of Vibrations and Waves* - Grade: A — 01/2018 - 02/2018

Personal projects

Blog writing — vitaliset.github.io
Vitali Set is my site about Machine Learning (posts in portuguese only for now). — 07/2020 - Now

Journal Club presentations — vitaliset/talks

- These talks are in portuguese only, but you can sneak-peek a english small presentation I did about **BorutaPy** clicking here. Some of the talks I like the most:
  - **Multiarmed bandits problem**: discussed how to approach the problem, both Bernoulli and non-Bernoulli, with drift or context, using epsilon-greedy, UCB and Thompson Sampling.
  - **Out of distribution generalization**: how to have a good performance across different environments even if you only have a few in your training data.
  - **Leaf Embedding**: using bagging of trees to find a distance between samples that focus on features that are relevant to the problem you are trying to solve and is scale-independent.
  - **Symbolic Regression**: we are able to create a population of functions and measure it's fitness using a performance metric using Darwin's idea of natural selection (survival of the fitest) we can refine the individuals to get a good estimator with explicit expression.

Open Source Software

Data Umbrella's PyMC Open Source Sprint July 2022 — pymc-devs/pymc

- PyMC is a Bayesian Inference Python library that uses Monte Carlo Markov Chains to sample from posteriori distributions. During this sprint I got to work with my Experian DataLab friend @pibieta on depecration of functions, docstring and minor changes while learning the fundamentals of a Open Source Software community.

Undergraduate Research Fellowships

- **FAPESP (São Paulo Research Foundation)**:
  Wada Property in Doubly Transient Chaos — 11/2018 - 09/2019
  Advisor: Prof. Rafael Ribeiro Dias Vilela de Oliveira
- **FAPESP (São Paulo Research Foundation)**:
  The Euler-Maclaurin formula and a few applications in Pure and Applied Mathematics — 09/2017 - 08/2018
  Advisor: Prof. André Pierro de Camargo
- **CNPQ (Brazilian National Council for Research and Technological Development)**:
  Numerical study of orbits in the 3-body problem — 08/2016 - 07/2017
  Advisor: Prof. Cecilia Bertoni Martha Hadler Chirenti
- **UFABC, Researching Since the First Day Program**:
  Introduction to stellar structure and evolution — 08/2015 - 07/2016
  Advisor: Prof. Cecilia Bertoni Martha Hadler Chirenti