Carlo Domenico Longo de Lemos

Senior Data Scientist @ Experian DataLab LatAm — São Paulo, Brazil

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

WORK EXPERIENCE

Serasa Experian (credit bureau) | Latin America DataLab — **São Paulo, Brazil**

Senior Data Scientist — 1y2mo - **05/2022 - Present**

- Currently leading development of large language models for Experian's legal team with ChatGPT-like human interactivity, enhancing internal legal query handling and operational efficiency.
- Pioneered DataLab scientist role in non-life insurance, establishing a framework for claims forecasting models, leading to a 50% performance increase over Experian's previous models. My MLOps template with Docker containers and Airflow DAGs for client backtesting automation reduced the manual task to a click of a button.

Data Scientist II — 6mo - **11/2021 - 04/2022**

- As the team lead for a group of junior data scientists within the AgriScore team, I spearheaded the development of a credit score targeting the agricultural sector. Our responsibilities encompassed:
  - Defining the appropriate farmer population for model development;
  - Streamlining the out-of-time cross-validation framework;
  - Constructing business-centric metrics for hyperparameter optimization.

  Our strategic use of Experian's internal data and agri-focused feature selection replaced an external data source that accounted for 21% feature importance in the previous model, enhancing autonomy while keeping the same performance.

Itaú Unibanco (bank) | Advanced Analytical Consulting Management — **São Paulo, Brazil**

Data Scientist I — 1y - **11/2020 - 10/2021**

- Utilized the loan team's experimentation for pricing studies, crafting a causal inference model using Causal Inference models to assess individual elasticity of the entire Itaú portfolio. This initiative introduced a new pricing dimension alongside risk default and impacted over 30 million customers.
- Led a data science team to design a model determining monthly sales targets for Itaú's managers, identifying product growth opportunities. An A/B test substantiated a projected annual revenue increase of 3% for tested products.

Data Science Intern — 1y1mo - **10/2019 - 10/2020**

- Enhanced business data processes using Python's pandas and SQL data querying, reducing the operation time from 24 hours to just 5 minutes.
- Developed a Machine Learning model to forecast client complaint probability in response to overdraft price hikes.

EDUCATION

University of São Paulo (USP) — **08/2021 - Present**

Master's degree in Probability and Statistics - *Part time student* — **GPA: 4/4**

Federal University of ABC (UFABC) — **05/2015 - 04/2021**

Bachelor's degree in Science and Technology (Major in Pure Mathematics) — **GPA: 3.88/4**

TOOLS

Statistics and Machine Learning - A/B Testing, Classification, Regression, Causal Inference, Natural Language Processing, Anomaly Detection, Clustering, Multi-armed bandits.

Python - Functional programming, Object-oriented programming, numpy, joblib, functools, returns, toolz etc.

Data Science Frameworks - sklearn, lightgbm, pytorch, tensorflow, hyperopt, skopt, imblearn, category_encoders, xgboost, langchain, time-robust-forest, statsmodels, nltk, river, aif360 etc.

Data Manipulation - pandas, PySpark, SQL, HiveQL, SASpy, Hadoop.

Data Communication - Strong data storytelling and visualization with Python's matplotlib, plotly and PowerBI.

Software - Git, VSCode, Bash, pytest, flake8, black, GitHub Actions, Airflow, Docker.

OPEN SOURCE SOFTWARE

scikit-learn/scikit-learn — **All my pull requests to scikit-learn**

- **ENH Adds support to sample weight to Partial Dependence plots - #25209 and #26644** (merged) +340 −28
  As partial dependence of a model is defined as an expectation, it should respect sample_weight if someone wishes to use Radon-Nikodym derivative to do importance weighting.
- **FEA Implementation of "threshold-dependent metric per threshold value" curve - #25639** (review) +304 −0
  Using 0.5 as threshold is suboptimal for many problems. This curve can help you choose a better threshold for your binary classification estimator. Related to my blog post.
- **ENH Add sample_weight parameter to OneHotEncoder's .fit - #26330** (review) +222 −23
  OHE offers the flexibility to exclude low-count categories, if desired. However, it does not take into account the samples' weight during the exclusion. This PR introduces the necessary logic to incorporate this functionality.

RESEARCH FELLOWSHIPS

FAPESP (São Paulo Research Foundation) — **11/2018 - 09/2019**
Wada Property in Doubly Transient Chaos

FAPESP (São Paulo Research Foundation) — **09/2017 - 08/2018**
The Euler-Maclaurin formula and a few applications in Pure and Applied Mathematics

CNPQ (Brazilian National Council for Research and Technological Development) — **08/2016 - 07/2017**
Numerical study of orbits in the 3-body problem

UFABC, Researching Since the First Day Program — **08/2015 - 07/2016**
Introduction to stellar structure and evolution

OTHER PROJECTS

Vitali Set — **vitaliset.github.io**

- Vitali Set is my personal site about Machine Learning. I'm starting to put videos of the posts on youtube.

Journal Club and professional training — **vitaliset/talks**

- I have a passion for continuous learning and knowledge sharing, so I regularly participate in and present at journal clubs and professional training, enriching both personal expertise and collective team knowledge.