Carlo Domenico Longo de Lemos

Data Scientist @ Experian DataLab LatAm

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

As a Machine Learning specialist, I have led end-to-end projects in the financial sector involving predictive and prescriptive analysis using ML. In addition to my advanced knowledge of probability and statistics, I'm a self-taught Python programmer with excellent computational skills following software engineering best practices. I don't mind being wrong and I always promote a safe environment, where *there are no silly questions*, valuing collaboration and the spread of knowledge among peers.

WORK EXPERIENCE

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

Senior Data Scientist **05/2022 - Present**

- I led the development of a Causal Inference model that estimates how the credit default probability of a person changes if we increase his credit card usage. The model was created using covariate adjustment (or S-learner, using causalml meta-learner jargon) and can be used to determine an individual's best credit card limit.
- I facilitated the life insurance team to apply domain adaptation techniques (importance weighting) to correct inconsistencies in the development dataset of the life risk classification model. This technique was responsible for a 5-point gain of ROC AUC in our lightgbm.LGBMClassifier model.
- I am currently working on the regression problem of predicting the severity of theft and robbery claims for insurance policies. My model outperforms Experian's generic benchmarks by over 30% using alternative data sources and a data-centric approach for constructing the development dataset.

Data Scientist II **11/2021 - 04/2022**

- I led a small team of junior data scientists in developing a Machine Learning model within the AgriScore team. We were responsible for building and maintaining a lightgbm.LGBMClassifier credit score aimed at the agricultural market. Activities included:
  - selecting the correct population of farmers for model development;
  - restructuring the out-of-time cross-validation framework;
  - designing business-driven metrics for hyperparameter optimization.

  Our team replaced an external data source responsible for 21% of the shap.explainers.Tree.shap_values feature importance of the previous model using Experian internal data and agri-oriented feature selection and engineering.

Itaú Unibanco (bank) | Advanced Analytics Products Management **São Paulo, Brazil**

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

- We took advantage of the RCT made by the loan team for pricing studies and developed a causal inference model (matching from sklearn.neighbors.NearestNeighbors of a Leaf Embedding from a sklearn.ensemble.RandomForestClassifier) to study the individual elasticity (CATE) of the entire Itaú portfolio. With this project, we created a new pricing dimension in addition to the risk of default in an easy-to-replicate way for other Itaú products impacting more than 30 million customers.
- Applying the same strategy of the loan pricing model, I led a small team of data scientists to develop a model that suggests Itaú's managers' goals (how much of each product they should sell that month). From the search

for similar managers, we were able to spot growth opportunities in certain products and suggest those changes. Based on an A/B test, the client area estimated the project's earnings at BRL 452 million per year (3% increase).

Data Science Intern 10/2019 - 10/2020

- I was responsible for topic modeling the texts from transcribed calls with a gensim.models.ldamodel.
- I automated some business manual data processes using Python's pandas and HiveQL for querying data with Hadoop.
- I created a sklearn.ensemble.RandomForestClassifier model to predict how likely the client will complain if we have an increase in overdraft price.

EDUCATION

Master's degree in Probability and Statistics | University of São Paulo (USP) - GPA: 4/4 **08/2021 - Present**

Bachelor's degree in Pure Mathematics | ABC University (UFABC) - GPA: 3.88/4 **05/2015 - 04/2023**

Bachelor's degree in Science and Technology | ABC University (UFABC) - GPA: 3.88/4 **05/2015 - 04/2021**

TOOLS

Machine Learning	Classification, Regression, Causal Inference, Anomaly Detection, Clustering, Natural Language Processing.
Python	Good understanding of object-oriented programming and SOLID principles as well as initial knowledge of functional programming. matplotlib, pandas, numpy, abc, collections, itertools, functools, toolz etc.
ML/DS Frameworks	sklearn, lgbm, pytorch, nltk, hyperopt, skopt, imblearn, category_encoders, xgboost, time-robust-forest, causalml, statsmodels, stable_baselines3, river, aif360 etc.
Database querying	SQL, HiveQL, SASpy, Hadoop.
Software	Git, GitHub Actions, Airflow, Docker, VSCode.
Languages	Portuguese (native) and English (advanced).

OPEN SOURCE SOFTWARE

scikit-learn/scikit-learn

- **ENH Adds support to sample weight to Partial Dependence plots #25209** (review) +172 −18
  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 Nearest neighbors search now allows for NaN euclidean metric #25330** (review) +154 −11
  Despite NaN euclidean being a scikit-learn metric, previous to this PR it was not possible to use it with nearest neighbors search because of input checks.

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 site about Machine Learning (only one post in english for now).

Journal Club presentations **vitaliset/talks**

- Some of the talks I like the most:
  - Multiarmed bandits problems: 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: on how to perform well 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 relevant to the problem you are trying to solve and are scale-independent.
  - Symbolic Regression: we can create a population of functions and measure their fitness using a performance metric. Using Darwin's idea of natural selection (survival of the fittest), we can refine the individuals to get a good estimator with explicit expression.

Imbalanced Binary Classification: A Survey with code **pibieta/imbalanced_learning**

- Together with my Experian DataLab friends Alessandro and Pablo, we wrote some chapters on approaching imbalanced binary classification. In particular, I discussed model interpretability under these scenarios.

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.