Carlo Domenico Longo de Lemos

Data Scientist @ Experian DataLab LatAm

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

As a Machine Learning proficient, I have led end-to-end projects in the financial sector involving predictive and prescriptive analysis. In addition to my advanced knowledge of probability and statistics, I'm a self-taught Python programmer with great computational skills following software engineering best practices. In all my professional experiences, I always promote the sharing of knowledge among peers: whether through the organization and presentation of journal clubs, as well as structuring internal training, study groups, and tutoring juniors and interns.

WORK EXPERIENCE

Experian (credit bureau) | **Senior Data Scientist** **05/2022 - Present**

- Led the development of a Causal Inference model that estimates the credit default sensibility of a person 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 the best credit card limit.
- I helped the insurance team apply domain adaptation techniques (importance weighting) to correct inconsistencies in the development dataset in the life risk classification model. This technique was responsible for a 5 point gain of ROC AUC in our lightgbm.LGBMClassifier model.
- Currently working on the regression problem of predicting the severity of theft and robbery claims for insurance policies. My model outperforms Experian's benchmarks by over 30% by the usage of alternative data sources and data-centric approach on constructing the development dataset.

Experian (credit bureau) | **Data Scientist II** **11/2021 - 04/2022**

- Led a small team of junior data scientists on Machine Learning model development within the AgriScore team. We were responsible for building and maintaining a lightgbm.LGBMClassifier credit score aimed at the agricultural market. Activities include selecting the correct population of farmers for model development, restructuring the out-of-time cross-validation framework and designing business-driven metrics for hyperparameter optimization. Our team was able to replace 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) | **Data Scientist I** **11/2020 - 10/2021**

- We took advantage of the RCT made by the loan team for pricing study, and were able to develop 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 dimension of pricing in addition to risk of default in an easy to replicate way for other Itaú products impacting more than 30 millions customers.
- Applying the same strategy used in the loan pricing model, I led the small team of data scientists in the technical development of a model that suggests the Itaú's managers' goals. From the search for similar managers, we were able to spot growth opportunities in certain products and suggest those changes.

Itaú Unibanco (bank) | **Data Science Intern** **10/2019 - 10/2020**

- I was responsible for topic modeling the texts from transcribed calls with a gensim.models.ldamodel.

- Automated business manual data process using Python's pandas and HiveQL for querying data with Hadoop.
- Created a sklearn.ensemble.RandomForestClassifier model to predict how likely the client is to complain if we have an increase in price of overdraft.

EDUCATION

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

Bachelor's degree in Pure Mathematics | **Federal University of ABC (UFABC)** — **05/2015 - Present**

Bachelor's degree in Science and Technology | **Federal University of ABC (UFABC)** — **05/2015 - 04/2021**

OPEN SOURCE SOFTWARE

Data Umbrella's PyMC Open Source Sprint — **07/2022**

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

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, statsmodels, tensorflow, nltk, hyperopt, skopt, imblearn, category_encoders, lgbm, xgboost, time-robust-forest, causalml, stable_baselines3, river, aif360 etc.
Database querying	SQL, HiveQL, SASpy, Hadoop.
Backend / Production	Git, Airflow, Docker.
Languages	Portuguese (native) and English (advanced).

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

PERSONAL PROJECTS

Vitali Set **vitaliset.github.io**

- Vitali Set is my site about Machine Learning (posts in portuguese only for now).

Journal Club presentations **vitaliset/talks**

- These talks are in portuguese only, but you can sneak-peek a small english presentation I did about BorutaPy. 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 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 its 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.

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.