Clébio Marques de Oliveira Júnior

Email: clebiomojunior@gmail.com
LinkedIn: https://www.linkedin.com/in/clebiojunior
GitHub: https://www.github.com/juniorcl

Professional Summary

Data Scientist with over 4 years of experience in machine learning, natural language processing (NLP), and data analysis, with a strong focus on extracting and processing unstructured data. Proven experience in credit scoring, sales forecasting, customer segmentation, churn prediction, and behavioral analysis. Skilled in building end-to-end data pipelines using Python, pandas, scikit-learn, spaCy, OCR, regular expressions, and FAISS.

Work Experience

Data Scientist – Vert Analytics

October 2024 – Present

- Automated legal defense recommendation systems using NLP techniques, embeddings, and FAISS.
- Built data extraction pipelines from PDF documents using OCR and regular expressions.
- Developed text analysis pipelines for social media data using spaCy, including tokenization, lemmatization, stopword removal, and grammatical correction.

Data Scientist – Datarisk

January 2022 – August 2024

- Developed a credit scoring model to support credit approval decisions, improving risk assessment and reducing default rates.
- Built a sales forecasting regression model, reducing Mean Absolute Error (MAE) by 12%.
- Performed customer segmentation using clustering algorithms to support targeted marketing campaigns.
- Implemented behavioral predictive models to identify churn propensity, credit risk, and employment instability.

Data Scientist – Be.X!

March 2021 – January 2022

- Extracted and structured unstructured textual data using regular expressions and natural language processing techniques.
- Designed rule-based anomaly detection models aligned with business requirements.
- Developed delivery delay prediction models using ranking-based evaluation metrics.

Education

Master of Science in Natural Sciences (Environmental Sciences)
Universidade Estadual do Norte Fluminense (UENF)
2017 – 2019

Bachelor's Degree in Physics (Teaching Degree)
Instituto Federal Fluminense
2010 – 2016

Certifications

- Statistics for Data Science and Machine Learning
- Text Summarization with Natural Language Processing
- Clustering: Extracting Patterns from Data
- Text Recognition with OCR and Python
- Machine Learning and Data Science with Python: From Zero to Advanced

---

Technical Skills

Programming Languages: Python, SQL
Libraries and Frameworks: pandas, scikit-learn, spaCy, FAISS
Machine Learning: regression, classification, clustering, time series analysis
Natural Language Processing: tokenization, lemmatization, sentiment analysis, text summarization
Data Processing: OCR, regular expressions, PDF and image data extraction
Tools: Git, Jupyter Notebook, SAS

Languages

Portuguese: Native
English: Professional Working Proficiency
Esperanto: Professional Working Proficiency