

Clébio Marques de Oliveira Júnior

clebiomojunior@gmail.com | medium.com/@juniorcl | linkedin/in/clebiojunior | github.com/juniorcl

Professional Summary

Data Scientist with **4+ years of experience** applying **machine learning, NLP, and data analysis** to solve real-world business problems with measurable impact.

Proven expertise in **credit scoring, sales forecasting, customer segmentation, and behavioral modeling**, as well as in the extraction and processing of **unstructured data** using advanced OCR and regex techniques.

Professional background in technology and consulting companies (Vert Analytics, Datarisk, Be.X!), delivering projects that range from **data exploration and predictive modeling** to **interactive dashboards in SAS**. Active portfolio on GitHub, technical articles on Medium, and recognized contributions on Kaggle.

Work Experience

Data Scientist – *Vert Analytics*

Oct 2024 – Present

- Designed strategic dashboards in **SAS** to support data-driven decision-making.
- Developed a robust pipeline for **unstructured data extraction** (PDF, OCR, regex), data cleaning, and time-series analysis.
- Built an **NLP solution for social media comment analysis** using **spaCy** (tokenization, lemmatization, stopwords removal, grammar correction), integrating insights into interactive dashboards.

Data Scientist – *Datarisk*

Jan 2022 – Aug 2024

- Built a **credit scoring model** with higher accuracy for credit approval, reducing default risks.
- Designed a **sales forecasting regression model**, reducing MAE by 12% and supporting strategic planning.
- Conducted **customer segmentation using clustering**, enabling targeted marketing campaigns and personalized strategies.
- Implemented **behavioral predictive models** to identify churn propensity, credit risk, and job instability, driving retention and risk reduction initiatives.

Data Scientist – *Be.X!*

Mar 2021 – Jan 2022

- Extracted and structured **unstructured text data** with regex and NLP techniques.
- Created **business-rule-based anomaly detection models** to identify outliers and mitigate operational risks.

- Developed a **delivery delay prediction model** using ranking metrics, enabling proactive logistics management.

Education

- **M.Sc. in Natural Sciences (Environmental Sciences)** – UENF, 2017-2019
- **B.Sc. in Physics (Teaching License)** – Instituto Federal Fluminense, 2010-2016

Certifications

- Statistics for Data Science and Machine Learning
- Text Summarization with NLP
- Clustering: Extracting Patterns from Data
- Text Recognition with OCR and Python
- Machine Learning and Data Science with Python A-Z

Technical Skills

- **Languages & Tools:** Python (pandas, scikit-learn, spaCy, pdfplumber, pytesseract), SQL, Git, SAS, Jupyter
- **Machine Learning:** regression, classification, clustering, neural networks, time-series forecasting
- **NLP:** tokenization, lemmatization, regex, summarization, sentiment analysis
- **Other:** OCR, interactive dashboards, model deployment, data visualization

Languages

- Portuguese – Native
- English – Full Professional
- Esperanto – Full Professional