Vitor Pinheiro

AI/ML Software Engineer
Manaus, Brazil (UTC-4, remote) | joaovpro17@gmail.com | linkedin.com/in/jvpro | github.com/jvitorcosta

SUMMARY

AI/ML engineer with 5 years shipping ML, LLM and data systems in fintech (lending, credit risk, fraud) and energy forecasting. Recent work includes a LightGBM fraud model that cut false-positive alerts by 99% for a US bank, an LLM pipeline that explains each fraud alert to analysts in plain language, and a LangGraph support agent that resolves 80% of internal tickets. Strong in Python, Spark and Databricks end to end, from data pipelines and MLOps to model serving and agents. Based in Brazil in UTC-4, the same time zone as US Eastern for most of the year, and open to remote work.

SKILLS

- **Languages:** Python, Scala, SQL, Clojure, JavaScript/TypeScript, HTML, CSS
- **Machine learning:** LightGBM, XGBoost, SHAP, Keras, TensorFlow, OR-Tools, time-series forecasting, experimentation
- **LLMs and agents:** LangChain, LangGraph, OpenAI, Mosaic AI Model Serving, LLM evaluation, Claude, Cursor
- **Data and MLOps:** Spark/PySpark, Databricks, Delta Live Tables, Unity Catalog, MLflow, Airflow, Streamlit, FastAPI
- **Cloud and DevOps:** Azure, Docker, Kubernetes, GitHub Actions (CI/CD), observability
- **Domains:** credit and lending (NPV, offer experimentation, policy datasets), fraud detection and transaction monitoring, time-series forecasting, LLM agents, MLOps

EXPERIENCE

Qubika · Senior AI/ML Engineer (client: Coastal Community Bank, US)
Remote, part-time contract · Feb 2026 – Jul 2026

- Productionized a LightGBM fraud detection model that, on a 24-month backtest, cut transaction-monitoring alert volume by 99.2% (12.9M to 98.5K) and lifted analyst-facing precision from 1.29% to 76.4%, a 131x reduction in review burden while still surfacing 45% of confirmed fraud.
- Took analyst load from about 538K alerts a month at 78 alerts per confirmed fraud to about 4.1K alerts a month at 1.3 alerts per confirmed fraud, a 60x efficiency gain.
- Designed and shipped the end-to-end Databricks platform behind the model: a Delta Live Tables pipeline materializing segment-specific feature tables in Unity Catalog, and a Databricks Asset Bundle orchestrating training, scoring and promotion as version-controlled jobs, with MLflow tracking and models served from the Unity Catalog Model Registry across environments.
- Built an explainability pipeline that turns each fraud alert into a reviewer-ready rationale by computing SHAP attributions and calling a Databricks-hosted foundation model through Mosaic AI Model Serving with structured prompts, producing customer- and transaction-context narratives that speed up case triage.

Nubank · Software Engineer (Machine Learning)
São Paulo, BR (hybrid) · Sep 2022 – Jul 2026 · promoted from Intern to Engineer

AI and LLM systems

- Built a Slack AI assistant with Python, LangChain, LangGraph and OpenAI over the framework's codebase, documentation and Slack channel history, running as a FastAPI service on Docker and Kubernetes. It resolves 80% of support tickets without a human (10 to 2 a week), freeing engineering time for infrastructure and features.
- Prompted, tuned and evaluated Databricks Genie as a self-service analytics assistant for the Lending business unit, defining business concepts and field-level metadata over core customer and loan datasets. The proof of concept with business users handled 3 of 5 test scenarios and surfaced the ETL and documentation gaps that shaped the adoption strategy.

Machine learning and experimentation

- Translated the Lending NPV framework from Scala to Python/PySpark, unifying profitability analysis across real-time credit models (loan eligibility and real-time valuation) with distributed and single-node simulations of loan performance across 50+ product lines.
- Co-designed Scala/Spark experimentation logic that randomized personal-loan offer parameters (installments, maximum principal, interest rates) on customer subsamples, reducing survivorship bias and producing 20%+ cleaner training data for credit policy decisions.
- Engineered critical features for credit risk models, improving performance through systematic feature-importance analysis and domain-driven transformations.

Data platform

- Reduced landing time for critical Lending core policy datasets by 34 hours and improved model feature freshness by 2 days by reordering ETL execution, pushing filters earlier in the Scala/Spark DAG and applying time-based partitioning.
- Co-developed a Scala data transformation framework with segment, metric and aggregation catalogs and a shared API, centralizing Lending business definitions, and automated .jar deployment to Databricks clusters with GitHub Actions.
- Led data governance initiatives for Lending, including the subdomain migration, SCR monitoring infrastructure and credit metrics catalog consolidation, moving 100% of the team onto the new infrastructure.
- Implemented data validation and monitoring pipelines that kept lending data reliable across systems.
- Implemented observability for the team's ML services and optimized CI/CD pipelines with GitHub Actions.

- Provided 24/7 support for the monitoring infrastructure, onboarding multiple business units with zero critical incidents.

Analytics tools

- Turned a ~2-day manual reporting process into a weekly task of under an hour by co-developing an internal low-code dashboard framework in Python/Streamlit, with reusable APIs for plots, filters and aggregations, S3 data loading, SQLite caching, CLI setup and python-pptx exports. Adoption grew from Lending to 19+ business units, and several teams replaced Looker at about 1 USD a day in infrastructure costs.
- Built analytics abstractions (filters, aggregations, re-aggregated monitorings) for the framework, cutting execution time from ~20 minutes to ~5 seconds, and a weekly KPI monitoring routine that standardized metrics across credit programs, which 12 business units used to automate their monitoring.
- Designed APIs over datasets and pre-aggregated monitoring data that let non-technical users build dashboards without code in a no-code dashboard builder MVP.
- As an intern, supported credit-risk and loan-performance model development, ran EDA on lending datasets, and wrote documentation and playbooks for ML workflows.

EloGroup · Data Scientist Intern (client: Comgás)
São Paulo, BR (remote) · Jun 2021 – Sep 2022

- Co-implemented a LightGBM/SARIMAX/Holt-Winters ensemble over 24 city-level demand series for Comgás, Brazil's largest natural-gas distributor, after running the EDA and ADF, KPSS, Granger and Ljung-Box tests. The forecast kept 18 of 24 citygates within the 5% fine threshold in demand programming with Petrobras and cut contractual fines by roughly R$200M in the first three months.
- Co-designed end-to-end ML pipelines in Azure ML with MLflow tracking, migrating a Databricks-only solution to a managed setup orchestrated by Azure Data Factory, from Azure Data Lake to Synapse views and Power BI dashboards, with automated data quality checks in a medallion architecture.
- Built a Streamlit app that automated a Box-Jenkins-style forecasting workflow, from visualization to model evaluation, for business stakeholders.
- Standardized text processing across consulting projects by co-implementing an internal NLP framework in Streamlit with spaCy and NLTK.

PROJECTS

- **Car Cyberdeck** (github.com/jvitorcosta/car-cyberdeck), 2026: open-source firmware for two ESP32 touchscreens and a Garmin watch app: one deck shows Spotify as a Pokémon battle with synced lyrics, the other drives the car's LED lights by touch, Siri, an Xbox pad or the watch, all over the phone hotspot. Has CI and tagged releases.
- **SolvedOR** (github.com/jvitorcosta/or-solver-streamlit), bachelor thesis: a Streamlit app that teaches Operations Research. Students write models in textbook notation, OR-Tools solves them, and Plotly draws the feasible region of two-variable problems. Live at solvedor.streamlit.app.

EDUCATION

Bachelor of Industrial Engineering, Universidade Federal do Amazonas (UFAM), Manaus, BR · Jan 2019 – Sep 2023 · GPA 9.07/10

- Relevant coursework: Operations Research, Simulation, Numerical Analysis, Probability and Statistics, Time Series, Optimization, Stochastic Processes, Data Analysis.

CERTIFICATES

- Google Data Analytics Professional Certificate, Coursera · June 2023
- Information Technology Technician, Matias Machline Foundation · December 2016

LANGUAGES

Portuguese (native) · English (advanced, C1) · Mandarin (intermediate, B1)