Alexey Velásquez

Machine Learning Engineer
Medellín, Colombia | +57 300 343 6954 | alexeyvb135@gmail.com | github.com/Alexey0424 | linkedin.com/in/alexey-ve | alexey0424.github.io

Summary

- Machine learning engineer with 3+ years of experience building and deploying self-hosted ML and LLM systems: speech intelligence, schema-constrained LLM extraction, RAG and interpretable gradient boosting. Shipped production systems for two US real-estate companies and a 1.5-year structural health monitoring research project.

Technical Skills

- **Proficient in:** Python, SQL. **Experience with:** TypeScript, Bash, PowerShell. **Languages:** bilingual Spanish (native) and English.
- **ML & AI:** PyTorch, scikit-learn, LightGBM, XGBoost, Hugging Face Transformers, Whisper, pyannote, Ollama, llama.cpp, RAG with pgvector, SHAP, Optuna.
- **MLOps & platforms:** Docker, FastAPI, PostgreSQL, GitHub Actions, MLflow, Databricks, Azure, Cloudflare Workers, n8n, CUDA.
- **Methods:** schema-constrained LLM decoding, Model Context Protocol (MCP), event-driven architecture (webhooks), A/B testing, statistical testing.

Experience

Data Scientist, Luis Buys Homes LLC — Florida, USA (Remote) Jan 2026 – Present

- Engineered a remote MCP server (TypeScript, Cloudflare Workers, Google OAuth) exposing 12 role-gated, read-only CRM tools to Claude with server-side per-user data scoping and 46 unit tests; rolled out to all 11 CRM users, who now query live data in plain language.
- Built a KPI engine on Cloudflare Workers that computes the owner's sales KPIs from CRM, Slack and Meta data, guarded by 266 Worker and 28 Apps Script tests across 20 audited deploys, giving one verified source for every weekly report.
- Re-architected CRM automation (Follow Up Boss, JustCall, Slack, Google Forms) into 3 event-driven webhook pipelines, cutting ~43,000 idle polling operations/month to ~3 per event while routing every lead with zero manual touch.
- Added AI call reports (summary, score, topics, sentiment) to every logged call and containerized BERT sentiment analysis feeding PostgreSQL and Power BI dashboards; a single-step write design removed a production race condition.

Data Scientist, Cactus Home Offer — Florida, USA (Remote) Jun 2025 – Dec 2025

- Built an end-to-end speech-intelligence pipeline (Whisper ASR, speaker diarization, speech-emotion recognition, SHAP) over real-estate sales calls, cutting manual call review by up to 80% and measurably improving closer performance.
- Replaced rule-based KPI extraction with schema-constrained LLM extraction (objections, buying signals, rapport, call outcome) over mixed Spanish/English transcripts, removing brittle keyword logic.
- Served CRM-triggered real-time inference through n8n with Azure-hosted services and Databricks APIs to automate sales coaching.

Machine Learning Engineer, EIA University — Medellín, Colombia Aug 2023 – Aug 2025

- Trained LightGBM and XGBoost on modal features (Frequency Domain Decomposition) from 6,000+ field vibration records of a real 14 m wind turbine, reaching up to 99.5% damage-detection accuracy with SHAP explanations validated against the structure's physics.
- Benchmarked Random Forest, SVM, LightGBM and autoencoders under stratified 10-fold cross-validation and designed PostgreSQL ingestion for continuous monitoring across a 1.5-year field campaign.
- Reproduced a published deep residual network in PyTorch (CUDA, Optuna) on raw accelerometer waveforms; per-window normalization lifted detection of never-seen damage from 47% to 98% (manuscript under peer review).

Projects

- **ORCA call intelligence (open source):** self-hosted GPU stack (Whisper, pyannote, emotion2vec, Ollama, pgvector RAG, FastAPI, Docker) turning raw sales calls into coaching scorecards and a chat-searchable archive; ~95% retrieval quality at 20% of the embedding storage.
- **Blueprint-to-Revit AI pipeline:** computer vision (OpenCV, PyMuPDF), parallel Claude perception agents, a custom MCP server and pyRevit automation that trace hand-drawn house plans into Revit 2024 sheets; 75 models delivered in 17 batches, agent turns cut by up to 72%.
- **LLM web scraper (open source):** local LLM plans the crawl and extracts records with schema-constrained decoding, verifying every value against the source page before export; 151 offline tests and a per-field precision/recall benchmark.

Education

Master of Engineering, Machine Learning focus, EIA University — 1st Class (94%) Jul 2025

Bachelor of Civil Engineering, Universidad de Medellín — Licensed engineer (COPNIA) Mar 2023