Erick Martinez

efm.physics@gmail.com | linkedin.com/in/erick-f-martinez | 818 292 4554 | Los Angeles, CA

Skills: Machine Learning (Random Forests, Clustering, Feature Engineering, Vector Search, Recommender Systems), Deep Learning (CNN, LSTMs, Transformers: BERT, RoBERTa, T5, GPT), NLP (Text Classification, Text Generation, RAG, Agentic AI), Computer Vision (YOLO, OpenCV), Object Detection, Image Classification, Data Analysis, Data Infrastructure, Data Visualization, Mathematics (Statistics, Linear Algebra, Multivariate Calculus), Data Ethics, Data Privacy, Algorithms

Platforms: Python, SQL, Git, Hugging Face, PyTorch, TensorFlow, scikit-learn, pandas, NumPy, Matplotlib, Plotly, Jupyter, Streamlit, LangChain, PySpark, Amazon Web Services (AWS), Google Cloud Platform (GCP), Snowflake, Jira, PyCharm

EXPERIENCE

TikTok

Data Scientist **Mar 2024 - Present**

- Developed high-precision **AI-Generated Content (AIGC)** detections, reducing false positives and improving model recall by **40%**. Designed a novel report-based **deepfake detection** and fine-tuned a **RoBERTa language model** using comment signals, increasing detection precision from **35% to 65%**—TikTok's most accurate AIGC detection method. **Actioned 25,000+ misleading AIGC videos**, tackling fraud, deceptive edits, and synthetic media abuse.
- Led investigations into sexual solicitation and abuse, actioning **50,000+ accounts** with **network-based** detection and **embedding search** using cosine similarity. Contributed to enforcement-focused policy updates.
- Developed a scalable **domestic terrorism** detection framework, actioning 1,200+ extremist accounts with network-based methods, driving high-priority actions ahead of the **2024 U.S. presidential election**.
- Mentored new hires, conducted peer reviews, and onboarded analysts. Led **cross-functional** collaborations on algorithm training pipelines, ensuring detection scalability and team continuity during restructuring.

Analytica Consulting

Data Engineer **Aug 2023 - Mar 2024**

- Engineered a comprehensive data processing workflow for parsing hundreds of thousands of HL7-XML encoded messages in minutes, using **Snowflake** and **schemachange**, to support a wide range of infectious diseases, enhancing public health analysis capabilities and supporting impactful insights into patient health trends.
- Key contributor in the migration and enhancement of the CalRedie **database**, improving data accessibility and granularity for various teams, thereby facilitating advanced public health **data analytics** and decision-making.
- Innovated in **data efficiency** by implementing a logging workflow in Snowflake, using SQL and Python. This system significantly improved error tracking, **cost analysis**, and process optimization. Reducing the number of validation queries needed by 7x, setting a new standard in **data engineering** within a public health framework.
- Developed high-impact analytical data queries as a Snowflake expert, influencing public health strategies and enhancing disease surveillance understanding, demonstrating the direct societal impact of data-driven decisions.

PDT AI

AI Engineer **May 2023 - Aug 2023**

- Pioneered a novel **Large Action Model (LAM)** product for business data analysis, focusing on email-based queries and data integration, using Pipedream and **Google Cloud Platform** to enable intelligent, context-aware responses
- Created a 'Cooperative AI' system using specialized **ReAct framework agents**, significantly enhancing response efficiency with notable reductions in response times by 60% and dramatically **decreasing error rates by 10x**
- Contributed to the **open-source** project, codeinterpreterapi, optimizing its error handling and performance for enhanced AI data analysis and visualization, effectively addressing complex queries across various industries

Spectrum Labs

Data Scientist **Aug 2022 - Apr 2023**

- Implemented state-of-the-art models (**LSTMs, Transformers: XLM-RoBERTa**) to accurately detect 21 unique behaviors, including Hate Speech and Self Harm, enabling clients to classify over half a billion messages daily
- Created an **internal python package** that streamlined data sampling workflows by executing highly parallelized, **asynchronous querying** of multiple Snowflake tables, reducing latency from 1 hour to just 10 minutes

- Designed a user-friendly **Streamlit** web app, allowing teams to easily perform statistically validated cross-client sampling with customizable parameters, simplifying the data sampling process for all users
- Streamlined data operations workflows to integrate with external platforms (**AWS, Snowflake, Jira**), resulting in a **5x increase in efficiency and throughput**
- Reduced company labeling costs by 30% whilst achieving high precision classification models through implementation of proactive sample analysis tools for data processes
- Achieved parity by reverse engineering competitor determinations using **decision tree** models, allowing us to understand their determinations and map our solutions onto prospective client traffic

Data Science Intern **Oct 2021 - Aug 2022**

- Developed datasets on sensitive toxic data using **OpenCV** and **OCR (Tesseract)** for user report screenshots
- Enhanced behavior lexicon and labeled datasets for multiple behaviors, contributing to more accurate analysis

RepairPal

Data Science Intern **Jun 2021 - Sep 2021**

- Developed 93% accurate prediction of vehicle repair costs by developing an unsupervised classification model using **K-means clustering** and **Principal Component Analysis** to categorize repairs and maintenance work
- Developed maintenance cost estimator (**Linear Regression**) to project cost of owning used vehicles over time

Corning Incorporated

Process Tech Specialist **Jun 2018 - May 2020**

- Implemented **Tesseract** and **OpenCV** to develop an efficient **OCR** system, resulting in a 8% increase in yields
- Designed an automated optical inspection process with database capabilities using Python, Shell, and VBA

PROJECTS

Aurora — *Generative Data Visualization* **2023**

- Developed a web app enabling users to upload documents for automated data analysis and visualization, integrating ChatGPT API with LangChain to generate and execute Python code.
- Leveraged Pinecone for vector storage and Plotly for interactive visualizations, enhancing data exploration.

Neutrally — *Text-to-text bias neutralization* **2021**

- LLM (**T5**) fine-tuned to reduce occurrences of inappropriate subjectivity in text. Labeled dataset composed of 55,503 biased and neutralized sentence pairs generated from Wikipedia edits tagged for “neutral point of view”
- Achieves state of the art performance with a BLEU score of 94.08 for bias neutralization

---

EDUCATION

University of California, Berkeley — *M.S. in Data Science* **2020 - 2022**

University of California, Santa Barbara — *B.S. in Physics* **2013 - 2017**