Trevor Santiago

trevorjsantiago1@gmail.com | (209) 485-0240 | Sacramento, CA | https://github.com/tjsanti/

SUMMARY

Data Scientist with 4+ years of experience building and deploying production data pipelines and applied ML systems in cloud environments, with recent focus on retrieval-based systems and LLM-powered data workflows for production applications and new product development.

EXPERIENCE

ALLDATA — Data Scientist | Nov 2021 – Present

- Built and deployed an AI system using FAISS-based retrieval and a fine-tuned Flan-T5 model to normalize automotive cause descriptions, enabling scalable SME validation workflows.
- Reduced SME effort by ~50% by shifting from manual authoring to AI-assisted prefill with validation and light edits.
- Built production data pipelines with LLM-based enrichment of free-text fields, improving data quality and coverage for a newly launched product that generated 85+ subscriptions within the first 2 weeks post-launch (~$100+/mo each).
- Rebuilt an Excel macro-based text-matching workflow as a Python semantic retrieval pipeline, reducing processing time from hours to minutes for datasets of 30k–120k records.
- Acted as the primary point of contact for data and AI systems across teams, collaborating with product, UI, and API leads to define requirements, guide implementation, and present system designs to stakeholders including executives.
- Designed and implemented BigQuery/Dataform data models and pipelines integrating multiple data sources to power downstream search and API systems.

New York Mets — Data Science Intern | Jan 2021 – Aug 2021

- Built an outfield defensive alignment model using Python, Scikit-Learn, and XGBoost, integrating upstream models for hit-type classification and catch probability into a single decision workflow.
- Engineered features by constructing a multi-dimensional hit outcome distribution using SciPy to estimate landing probabilities and expected outcomes used by the alignment model.
- Developed Matplotlib evaluation and diagnostics (error analysis and performance tracking) to validate model behavior across scenarios and guide iteration.
- Created a custom dataset of MLB venue outfield wall distances using interpolation to generate structured spatial features for predictive modeling.

EDUCATION

University of San Francisco — M.S. Data Science | Aug 2021
Relevant coursework: Machine Learning, Deep Learning, SQL, Distributed Computing (Spark), Data Structures & Algorithms
University of California, Santa Barbara — B.S. Mathematical Sciences | Jun 2020
Minor: Statistical Science

SKILLS

Programming & Querying: Python, SQL
AI & Machine Learning: Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), LangChain/LangGraph, AI Agents, prompt engineering, Scikit-Learn
Data Engineering & Processing: ETL/ELT, Pandas, NumPy
Cloud & Deployment: Google Cloud Platform (BigQuery, Cloud Run, Cloud Functions, Cloud Storage), Docker
APIs & Backend: FastAPI, Flask