Isai Garcia-Baza

https://isaigb.github.io/ github.com/isaigb linkedin.com/in/isaigarciabaza

Research and Work Experience

U.S. National Science Foundation **Alexandria, VA**

Data Scientist *Sept. 2024 - Present*

- **Deployed production-grade NLP pipeline** leveraging Python, Hugging Face, AWS, and SQL to transform unstructured text into semantically enriched, model-ready datasets.
- **Built modular text classification framework** supporting model versioning, ensemble averaging, uncertainty quantification, and one-click reviewer dashboards **reducing human review workload by 40%**.
- **Evaluated/Fine-tuned embedding models** for classification evaluating separately for semantically distinct documents (ROC/AUC, recall).
- Developed and operationalized **ETL pipelines** for labeled data ingestion, ensuring reproducibility and governance compliance.
- Collaborated with **Data Engineering, Architecture, and cross-functional stakeholders** to improve internal data quality and reporting.
- Led quantitative analyses of applicant performance, presenting actionable insights to Assistant Director-level leadership to inform program strategy.
- Contributed to **AI, NLP, and LLM Community of Practice**, sharing reusable code and best practices for Responsible AI.

UNC-CH School of Education **Chapel Hill, NC**

Graduate Researcher *Aug. 2021 - Present*

- **Built LLM-powered image segmentation & classification pipelines** using OpenAI, Gemini, and Ollama; optimized prompt design to align outputs to human-labeled semantics and produced structured, analysis-ready datasets.
- **Engineered robust ETL system** to process 15M+ records with validation, quality checks, error recovery, and JSON/Excel metadata; enabled efficient downstream modeling and querying.
- Applied **data balancing techniques** (SMOTE, random over/under-sampling) and K-fold cross-validation to optimize performance on imbalanced datasets, achieving balanced accuracy improvements.
- Engineered **feature-rich model prototypes** through iterative cycles of stakeholder feedback, feature engineering, and performance tuning.
- Strengthened **causal inference validity through propensity score matching**, enabling more reliable program impact evaluations.
- Partnered with faculty, analysts, and policy stakeholders to translate complex statistical findings into actionable recommendations.
- **Designed and trained predictive models** (Random Forest, LASSO regression) on large administrative datasets (13M+ observations from 16 institutions), improving grade prediction accuracy and informing student support strategies.

Data and Programming Skills

Python: Intermediate. 4 years of experience. Hugging Face, OpenCV, SQLite3, scikit-learn, imbalanced-learn, statsmodels, pandas, NumPy, Ollama, NLTK, Gensim, PyTorch, PyMuPDF.

R: Intermediate. 3 years of experience, mainly for coursework and RMarkdown.

STATA: Advanced. 8+ years of experience. Developer of internal tools for data cleaning, management, and monitoring.

Reporting and Data Visualization: Streamlit, Jupyter Notebook, Matplotlib, R Markdown, ggplot2, Quarto.

Others: Git, GitHub, SQL (PostgreSQL, SQLite, DBeaver), Bash, Docker.

Statistical Expertise: Causal Inference, Statistical Modelling, Machine Learning, Econometrics.

Education

University of North Carolina at Chapel Hill **Chapel Hill, NC**

Ph.D. Education Policy, graduate minor in Computer Science *Expected 05/26*

- **Selected Coursework:** Causal Inference, Machine Learning, Natural Language Processing, Linear Regression, Time Series, Multilevel Modelling

University of North Carolina at Chapel Hill **Chapel Hill, NC**

B.A. Psychology with Honors *May 2017*