Isai Garcia-Baza

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

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, Panel Methods

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

B.A. Psychology with Honors — *May 2017*

Data and Programming Skills

Python: Intermediate. 2 years of experience. NLTK, Gensim, Hugging Face, scikit-learn, imbalanced-learn, statsmodels, pandas, NumPy, TensorFlow.

R: Beginner. 2 years of experience, mainly for coursework and RMarkdown.

STATA: Advanced. 6+ years of experience. Develop internal tools for data cleaning, management, and quality monitoring.

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

Others: Git, GitHub, SQL, Bash, Docker.

Data Collection: Survey (RedCap, SurveyMonkey, SurveyGizmo), focus groups, interviews in English, Spanish.

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

Research and Work Experience

U.S. National Science Foundation — Alexandria, VA

Data Scientist Trainee — *Sept. 2024 - Present*

- Developed NLP pipeline to streamline eligibility assessment, to reduce human-reviewer workload (Libs: `Hugging Face, Gensim, NLTK, Scikit-Learn` Models/Algos: `SciBERT, Doc2Vec, Multinomial Naive Bayes, Random Forest`)
  - ∗ Collected and cleaned text data from PDF and .txt files
  - ∗ Collected and cleaned decision data across various Excel files
  - ∗ Trained and evaluated Doc2Vec embedding model
  - ∗ Trained eligibility classifiers using Doc2Vec embeddings
  - ∗ Trained and evaluated Multinomial Naive Bayes classifier using tf-idf bag-of-words features
- Designed and executed analyses of applicant performance `SQL, Python`
- Created data visualizations for briefings (nontechnical audience)

UNC-CH School of Education — Chapel Hill, NC

Graduate Research Assistant — *Aug. 2021 - Present*

- NSF STEM (Github) – Python: `scikit-learn` , `imbalanced-learn` ; STATA
  - ∗ Analyze large-scale administrative data by first combining administrative data sourced from 5 databases across 16 institutions resulting in 13 million observation data set
  - ∗ Trained Random Forest and alternative classification models to predict student grades
  - ∗ Trained Random Forest, LASSO regression models for comparison against classification performance
  - ∗ Iterated over model prototyping, stakeholder input, and feature engineering cycles
  - ∗ Balanced data using SMOTE, Random Over Sampling, Random Under Sampling
  - ∗ Tuned models using K-fold cross validation and assessed using balanced accuracy to select best model
- Developed internal STATA tool for automatically generating data quality reports with variable labels
- Strengthened causal inference claims by implementing propensity score matching

Child Trends — Bethesda, MD

Senior Research Assistant — *Aug. 2017 - Jan. 2021*

- Webscraped 323k Tweets to analyze trends in social media conversations on AI/AN issues
- Analyzed national audience reach of television news segments providing insight for content optimization
- Created and analyzed randomized control trial surveys
- Drafted internal and external reports of research findings