Henry Caldera

661.340.2028 | henry@datascientist.ai | LinkedIn Profile

Education

Master of Science, Information and Data Science | University of California, Berkeley *May 2024*

Bachelor of Science, Biochemistry | University of California, Santa Barbara *Jun. 2015*

Skills

Programming Languages [libraries]:
- Python [TensorFlow, Keras, PyTorch, Scikit-learn, NLTK, SpaCY, Pandas, NumPy, FastAPI, Pytest, Pydantic, prometheus, k6, Grafana]
- R [Tidyverse, lmtest, sandwich, fable, feasts, plm]

Data & ML Engineering [databases]:
- SQL [PostgreSQL], NoSQL [Redis, MongoDB, Neo4j], Docker, Kubernetes, Git, Github

Machine Learning [architectures]:
- Large Language Models [encoder-only, encoder-decoder, and decoder-only transformers]
- Deep Neural Networks [multi-layer perceptron, recurrent neural networks, convolutional neural networks]
- Algorithms [k-means clustering, k-nearest neighbors, XGBoost, random forest]

Statistical Methods:
- Experimental design, field research, hypothesis testing [A/B, multivariate]
- Cross-sectional statistical modeling and inference [discrete response, continuous response]
- Time-series modeling and forecasting [ARIMA, VAR], and panel data analysis.

Data Visualization [libraries]:
- Python [Matplotlib, Seaborn, Plotly], R [ggplot2], Power BI

Cloud Computing Platforms:
- Amazon Web Services (AWS), Google Cloud Platform (GCP), Microsoft Azure

Projects

MedQuest: Your path to health, simplified. *Jan 2024 - Apr 2024*

An app created for Berkeley's Capstone project leveraging LLMs and deep learning modeling to enable disease classification from symptoms using natural language input. With a mission to bridge the gap in accessibility between those who need healthcare and those who can afford it, our platform is designed to provide quick assessments and guidance on the next steps for their health concerns using natural language input; a first of its kind.

Investigating Neural Machine Translation Strategies for Tagalog *Aug 2023 - Dec 2023*

Conducted LLM experiments to improve translation performance on low-resources languages by testing the effect of fine-tuning LLMs with varying percentages of mixed real-synthetic data (English-Tagalog sentence pairs) created using the back-translation technique. Decoder-only LLMs (ChatGPT-3.5) showed only marginal benefits from training on mixed real-synthetic datasets, while encoder-decoder LLMs (M2M100 & mBART50) showed significant increases in BLEU and BLEURT metrics, as well as native-speaker evaluations, which demonstrated improvements in fluency, adequacy, and formally of translated texts to the target language.

Time Series Forecast of the Keeling Curve *Aug 2023 - Dec 2023*

With Mauna Loa Observatory's data of atmospheric CO2 concentration from 1998 to 2023, a seasonally-adjusted ARIMA model was fit to the weekly Keeling curve data from 1998 to 2021. The SARIMA model's performance was evaluated comparing the predicted and realized CO2 concentrations between 2022-2023 and showed a normalized RMSE of 0.065. Critically, the SARIMA model forecast predicts the atmospheric CO2 concentration will reach the 500 ppm threshold in December 2061 - well within our lifetimes.

Henry Caldera

661.340.2028 | ecalderajr@berkeley.edu | Berkeley Grad Profile

Projects

Effect of Changing Traffic Laws on Rates of Traffic Fatalities Across States *Aug 2023 - Dec 2023*

With monthly panel data from the National Highway Traffic Safety Administration between 1980-2004 for the 48 continental U.S. states, modeled the effect of changing regulatory traffic laws on the total fatalities per 100,000 people in state population. Controlling for state-level fixed effects, the findings include a decrease of 1.1 deaths ($p<0.05$) where driver's BAC limits are 0.08, a decrease of 1.2 deaths ($p<0.01$) where per se DUI laws exist, a decrease of 1.2 deaths ($p<0.05$) where primary seatbelt laws exist, a decrease of 1.2 deaths ($p<0.01$) where graduated drivers license laws exist, an increase of 1.2 deaths ($p<0.01$) for every percentage point increase of the population aged 14-24, and a decrease of 1.4 deaths ($p<0.01$) for every one-point increase in the log of the unemployment rate - highlighting the importance of both traffic laws and population demographics on the effect of traffic fatality rates.

Effect of Additional Road Signs on Cyclist's Traffic Law Compliance *May 2023 - Jul 2023*

Conducted a single-blinded field experiment with clustered randomization and between-subjects design to test the effect of additional targeted signage at stop-controlled intersections on changes in cyclists' compliance with stopping and yielding. The treatment intervention increased the propensity of cyclists to abide by the traffic law and stop by 17% ($p<0.01$). Additionally, the intervention demonstrated an increase of 36% in the propensity of cyclists to yield ($p<0.01$).

Prediction of Age-related Disease from Patient Health Data *May 2023 - Jul 2023*

With data from InVitro Cell Research, a company focused on personalized regenerative medicine, created random forest, XGBoost, and deep neural network models to predict disease presence based on 56 patient characteristics - weighted F1 score of 90.0%.

Effect of PM 2.5 on Asthma-related visits to the ER in Los Angeles *Jan 2023 - Apr 2023*

With data from the California Environment Protection Agency, used OLS regression to model the effect of particular matter (<2.5 μm) on rate of Asthma-related ER visits in Los Angeles. The multivariate model, controlled for socioeconomic factors, showed an increase of 8 Asthma-related visits per 1 million ER visits for every one increase in percentile rank of PM_2.5 concentration ($p<0.01$). Importantly, the analysis revealed the impact of socioeconomic burden lead to an increase of 38 Asthma-related visits per 1 million ER visits ($p<0.01$), highlighting the outsized importance of social determinants of health in public health outcomes.

Experience

Electronic Health Record System Administrator | Threemile Canyon Farms *Apr. 2016 – current*

- Managing EHR system protocols to align with operational protocols by implementing new user functionalities and/or constraints in the web app and mobile app interfaces.
- Diagnosing problems with the EHR system and communicating directly with the vendor to report bugs and request new features.
- Ensuring the EHR system is connected for the import and export of data to external sources and systems, and controlling access to relational databases by managing EHR system accounts and their permissions.
- Generally oversee the entry, validity, and storage of 20+ million records.

Operations Data Analyst | Threemile Canyon Farms

- Generated ad hoc operational and performance reports through collaboration with cross functional groups such as management and accounting.
- Guided strategic decisions through data analyses and development of scorecards used by management teams to assess performance of all departments, resulting in (over 3 years):
    - 15% increase in production
    - 30% reduction in replacement costs
    - 40% reduction in mortality rate
    - 5% increase in reproductive performance

In Vitro Fertilization Ops Lead | Threemile Canyon Farms

- Gained leadership of early-stage pioneering bovine IVF program and developed protocols and standard operating procedures for the reproductive operations.
- Scaled operations over a two-year timeline by increasing business inputs i.e. expanding team size, source materials, equipment, and allocated land space to realize a 4x increase in monthly embryo production - culminating in the achievement of a then-world record.

Medical Research Assistant | Kern Medical Center *Mar. 2010 – Sept. 2012*

- Worked in the Emergency Department determining eligibility of prospective study participants, conferring with nurses and medical residents to verify criteria, and enrolling patients into medical studies.
- Secondary work in the microbiology lab preparing nasopharyngeal samples for genotypic analysis.
- Lead Assistant to Principal Investigators on study aimed to describe the spectrum of viruses in children presenting with bronchiolitis.
- Assisted on medical team's involvement in Macrogenics Inc.'s Phase II clinical trial of a vaccine for West Nile Virus in ED patients.

Publications

Walsh, Paul; Vieth, Teri; Rodriguez, Carolina; Lona, Nicole; Molina, Rogelio; Habebo, Emnet; Caldera, Enrique, et al. Using A Pacifier to Decrease Sudden Infant Death Syndrome: An Emergency Department Educational Intervention. *PeerJ* 2014; 2:e309.

Walsh, Paul; Overmyer, Christina; Hancock, Christine; Heffner, Jacquelin; Walker, Nicolas; Nguyen, Thienphuc; Shanholtzer, Lucas; Caldera, Enrique, et al. Is the Interpretation of Rapid Antigen Testing for Respiratory Syncytial Virus as Simple as Positive or Negative? *Emergency Medicine Journal* 2014; 31:153–159.

Walsh, Paul; Merchant, Sabrina; Aguilar, Valerie; Rodriguez, Carolina; Molina, Rogelio; Heffner, Jacquelin; Caldera, Enrique, et al. Performance of a Rule to Predict Apnea in Bronchiolitis. Abstract. *Society for Academic Emergency Medicine* 2011; 18(5):S85.
