Henry Caldera

661.340.2028 | henry@datascientist.ai | /in/henry-caldera

Education

Master of Science, *Information & Data Science* | University of California, Berkeley *Aug 2022 – May 2024*

Bachelor of Science, *Biochemistry* | University of California, Santa Barbara *Aug 2012 – Jun 2015*

Skills

Programming Languages [libraries]:

- Python [TensorFlow, Keras, PyTorch, Scikit-learn, Transformers, 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 /in/projects

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

A chatbot web app 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.

Projects

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

TA, Applied Machine Learning University of California, Berkeley | *Aug 2024 – current*

- Assisting graduate students with coding and concepts during office hours.
- Reviewing course map and proposing updates to teaching materials.
- Grading course assignments and projects.

Operations Data Analyst Threemile Canyon Farms | *Apr. 2016 – current*

- Generated ad hoc operational and performance reports for all departments across the operation 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

Electronic Health Record System Administrator

- 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 for support.
- 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.

In Vitro Fertilization Ops Lead

- 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.

Publications iD /0009-0001-7140-8447

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.