Thomas Lee

510-230-7619 | thomascl@berkeley.edu | https://thomascl24.github.io/ | linkedin.com/in/thomaslee24

Professional Profile

Early-career data professional with two years of experience in data science and machine learning, leveraging analytical thinking and technical proficiency to deliver impactful insights. Skilled in Python and machine learning with a strong foundation in statistical analysis for data-driven decision-making and a growing portfolio of applied data science projects.

Education

University of California, Berkeley — Berkeley, CA

Master of Information and Data Science | 4.0 GPA — August 2025

- **Relevant Coursework:** Machine Learning Systems Engineering (Docker, Kubernetes), Natural Language Processing with Deep Learning, Research Design and Applications for Data and Analysis, Statistics for Data Science

B.A. in Computer Science, Minor in Data Science | 3.9 GPA — May 2024

Experience

East Bay Municipal Utility District — Oakland, CA

Data Science Intern — May 2024 – Present

- Designed and deployed a fully tested Python module to streamline the processing of EBMUD's customer billing data for the Cost of Service (COS) study.
- Performed sensitivity analyses on rate structures to assess revenue impacts during low water consumption periods, reinforcing trust in rate structure stability with lower service charges.
- Analyzed customer billing data to answer key policy questions such as typical water usage patterns and the bill impacts of proposed rate structures, informing internal decision making and outreach efforts.
- Implemented a Python-based likelihood of failure (LOF) model to quantify the rate of pipe degradation, reducing a month-long QA process to less than a week and identifying critical data validity issues with RMIDs and missing leak data.
- Conducted water savings analysis for an advanced metering infrastructure (AMI) initiative, applying Bayesian inference to quantify savings distributions and inform long-term investment strategy.
- Streamlined AMI vendor data access by implementing an Amazon Redshift-based pipeline, replacing costly **$1,500** per request extractions with automated data streaming.
- Managed data access and privacy workflows for company collaborations with three UC Berkeley research groups, providing mentorship with data analysis and ensuring regulatory compliance.

Eikon Therapeutics — Hayward, CA

Machine Learning Intern — May 2022 – Aug 2023

- Identified **20** promising drug treatment candidates from over **200,000** compounds using an extended isolation forest machine learning model for anomaly detection.
- Created a neural network architecture with TensorFlow, classifying protein agonists and antagonists with over **84%** accuracy from a highly noisy dataset.
- Engineered a Python-based ETL data pipeline to accelerate data extraction, cleaning, and preprocessing, boosting data request efficiency by **22%** for a team of **10+** scientists.

Projects

RoBERTa and BERT Pokémon Trading Card Game (TCG) Card Power and Synergy Prediction — April 2025

- Enabled early price estimation for newly released cards by modeling latent card strength prior to observed gameplay performance and predicted novel card combinations with high synergy to give players a competitive edge.
- Fine-tuned BERT and RoBERTa-based NLP regression models to identify relevant information from raw card text.

San Francisco Crime Interactive Dashboard — March 2025

- Constructed interactive Tableau data visualizations for temporal and geographic analysis of crime in San Francisco.
- Created a Flask web application to embed data visualizations into a user-friendly dashboard.

Skills

Programming Languages: Python (Tensorflow, Keras, NumPyro, XGBoost, NumPy, Pandas, Matplotlib, Seaborn, Scikit-Learn), SQL (MySQL, Amazon Redshift), R, NoSQL (Neo4j)

Data Visualization and Analysis Tools: Tableau, Power BI, Dash

Other Tools: Kubernetes, Docker, AWS, Git/GitHub, HuggingFace, Visual Studio, RStudio