Veena Ramesh

veenaramesh.github.io | github.com/veenaramesh | linkedin.com/in/veenarameshh
+1 (703) 626 3378 | veena.ramesh.1@gmail.com

EDUCATION

University of Virginia **Charlottesville, VA**
Bachelor of Arts in Computer Science, Cognitive Science (Linguistics concentration) May 2021
Coursework: Machine Learning, Artificial Intelligence, Statistical Learning & Graphical Models, Natural Language Processing, Information Retrieval

EXPERIENCE

Lovelytics Data, Inc. **Washington, D.C.**
Senior Machine Learning Engineer (Level II) April 2022 – Jan 2024
Senior Machine Learning Engineer (Level III) Feb 2024 – Present

- A boutique data consulting firm that works largely with Databricks, receiving investment from both Databricks Ventures and Interlock Equity in 2023; won Databricks Innovation Partner of the Year (2022) and Databricks CME Partner of the Year (2023).
- Fine-tuned Llama 2.0 on an instruction dataset of campaign text messages using QLoRA and PEFT to increase writer productivity and produce novel messages; productionized using MLFlow and Mosaic AI Model Serving.
- Accelerated a Random Forest pipeline by migrating to the XGBoost algorithm, to take advantage of parallel processing on datasets; packaged and orchestrated using Apache Airflow and MLFlow for reproducible development and evaluation.
- Leading client-facing projects with large commercial (e.g. Rockstar) and federal clients (e.g. Center of Medical Services) as well as smaller business clients as well; communicating non-technically, and optimizing results while managing stakeholder expectations.

Itek Infomatic Inc. **Washington, D.C.**
Data Scientist June 2019 – March 2022

- Enhanced Random Forest model performance by leveraging Principal Component Analysis (PCA) for dimensionality reduction and a Random Forest algorithm for data imputation, improving testing accuracy by 3% (to 84%).
- Built a SQL ingestion pipeline to ingest and standardize data by adjusting unique weight specifications for downstream users.
- Analyzed differences between company and public data. Optimized queries by vectorizing and query tuning SQL statements.

Red Oak Strategic **Washington, D.C.**
Data Science Intern June 2019 – August 2019

- Red Oak is a start-up data science consulting company. Specialized in predictive modeling, business analytics, and AWS services.
- Engineered a time-series forecasting model (ARIMA), using AWS Sagemaker, to predict the likelihood of wildfires in California.

Columbia University Center of Justice **New York City, NY**
Justice Through Code Mentor Feb 2023 – June 2023

- Selected as a volunteer for JTC, which provided programming education in Python to those who have conviction histories.

PROJECTS

arXiv Search Engine May 2021, Feb 2024

- Built an end-to-end arXiv search engine by scraping the arXiv site, embedding the data using spaCy's built-in embedding model, chunking and vectorizing documents using the Gensim FastText model, and indexing data using the BM25 algorithm. Revisited the search engine to optimize embeddings using recent Large Language modeling techniques.
- Across approximately 60 surveyed students, the first iteration of the search engine performed qualitatively better than the existing arXiv search engine; consistent feedback showed that the engine retrieved more 'recent and relevant' research.

Analyzing Celebrity Language on Twitter (now X) May 2021

- Analyzed behavioral differences in political language on Twitter between celebrities (defined as verified users) and non-celebrities (identified as unverified users) using a transformer model and rule-based sentiment model (e.g. VADER) across a variety of topics, including COVID-19 (~95K users) and the primary presidential campaign (~900K users) during the 2020 election year.
- Demonstrated a statistically significant difference between the two populations through the Welch's unequal variances t-test. Verified users use more neutral sentiment than unverified users, who use more quantitatively extreme language.

SKILLS

- **Languages**: Python (advanced), SQL (intermediate; mySQL, DBSQL), Scala (medium; spark), R (medium; dplyr and ggplot).
- **Libraries and Frameworks:** PySpark, scikit-learn, mlx, PyTorch, spacy, boto3 (AWS).
- **Tools and Platforms**: AWS, Databricks, MLflow, Apache Airflow.

INTERESTS

- **Technical**: Generalized learning (i.e. like Meta's JEPA), RPG video game development, climate modeling, the mlx framework.
- **Personal**: Oil painting, learning real and fake languages, watching films, reading and writing poetry, playing video games.