ALICE HUA

Los Angeles, CA
626.677.1998 | alicehua11@berkeley.edu
linkedin.com/in/alicehua/
github.com/alicehua11
alicehuacal.com

EDUCATION

University of California, Berkeley Berkeley, CA
Master of Information and Data Science (4.0/4.0 GPA) Dec 2021
Relevant coursework:
Statistics for Data Science, Applied Machine Learning, Introduction to Data Engineering, Deep Learning on the Cloud and at the Edge, Experiments & Causality, Machine Learning at Scale

Bachelor of Arts in Geography (Magna Cum Laude) May 2020

TECHNICAL SKILLS

Python • SQL • Spark (PySpark) • R • Scikit-learn • Pytorch • LLM • Probability • ETL, • End-to-end ML systems • Classification/Clustering • Git • CI/CD • AWS • ETL • Airflow • Docker • Kubernetes • Terraform • Tableau • Snowflake

PROFESSIONAL EXPERIENCE

CrowdStrike, Sunnyvale CA Sep 2021 - Present
Data Scientist III – Detection June 2023 - Present
- Discover and train a binary classification model Linux model on Falcon telemetry data to detect malicious processes on over all customers with Falcon sensors on their Linux endpoints
- Develop and train the first behavioral model with big data to detect malicious file write using distributed Xgboost on Spark on Window endpoints which make up >80% of our customers
- Assess model performance and validation on customers' data collected using Logscale query language and processed with Python, AWS S3, Athena, EMR. Present models to stakeholders at model stage exit meetings
- Develop & design Active Learning Framework for ML models that suffer from data scarcity like Mac and Linux using querying strategies such as least confident, marginal difference, duplication using leaf nodes and specific filtering criteria
- Conduct POC for anomaly detection in partnership with NVIDIA using Morpheus and Falcon data
- Estimate model production cost, provide solutions to reduce volume and speed up prediction time for model funding

Machine Learning Platform Engineer II – ML Platform Aug 2022 – May 2023
- Lead interns and developed an automated ML Feature Store caching system upon deployment of new ML models using JIRA and Slack automation, AWS EMR and Apache Airflow, Snowflake
- Maintained production ML systems & in-house API services, internal infrastructure on Terraform (using Atlantis) alongside other engineering teams to ensure on time model delivery quarterly
- Collaborated with colleagues to build the pipeline for a whitelisting model with telemetry data aggregation done in Redshift and stored in Cassandra

Data Engineer Jan 2022 – July 2022
- Designed and built a Feature Store service supporting over 14 ML models in Apache Airflow, AWS Glue and Snowflake
- Integrated existing feature extraction service, AWS Batch, Athena, SQS to pull from Feature Store to reduce over $308k/year in AWS cost. Additionally, worked on a cost project of AWS S3 inventory to further reduce operational costs
- Built dynamic compute environment capability, parallelize Jenkins build, manage Snowflake data warehousing

Data Engineer Intern Sept 2021 – Dec 2021
- Built ETL pipeline in AWS Glue for our in-house feature extraction service to understand per iteration system metrics & metadata. Visualized results in Superset dashboards.
- Stood up MLFlow sever in production Kubernetes infra for daily model versioning and experiment logging

PAPERS & PROJECTS

Computer Vision & Edge Deployment Peer Reviewed Research Paper: Protecting endangered megafauna through AI analysis of drone images in a low-connectivity setting: a case study from Namibia – Nonprofit organization Wildtrack partnership
NLP Modeling with GPT-3: MemorAI – NLP Models of Loved Ones – Master of Data Science capstone project at Berkeley

CERTIFICATIONS

AWS Certified Cloud Practitioner & Developer Associate & Machine Learning Specialty Mar 2022 - Present