Sourabh Swargam

swargam.sourabh@gmail.com | LinkedIn

Professional Overview: Results-driven Data Scientist with 4+ years of experience developing scalable ML models, data pipelines, and cloud-based solutions across healthcare and analytics. Proficient in Python, SQL, AWS, and TensorFlow, with a strong foundation in statistical modeling, machine learning, anomaly detection, and A/B testing. Experienced in building automated data processing systems, optimizing ML workflows, and deploying predictive models for large-scale datasets. Enthusiastic about leveraging data science to drive business impact, optimize decision-making, and create innovative solutions.

Skills

Languages: Python, Java, JavaScript, HTML/CSS, SQL, R
Frameworks: TensorFlow, Keras, PyTorch, LLMs, Pandas, NumPy, Matplotlib, Seaborn, Flask, FastAPI, Hugging Face, Spring Boot, ReactJS
Tools: Docker, Kubernetes, Git, AWS (Beanstalk, Lambda, Step Functions), MySQL, MongoDB, Postman, Tableau

Experience

Data Scientist Oct. 2024 – Present
National Institutes of Health (NIH/NIDA) *Baltimore, MD*

- Applied machine learning and statistical modeling to large-scale public health and clinical datasets, ensuring compliance with HIPAA guidelines and data governance policies
- Designed and implemented a secure AI-driven chatbot application enabling peer coach–mentee conversations for research participants, facilitating qualitative data collection for clinical and behavioral studies
- Developed backend using Django integrated with LLM APIs for natural language processing and conversation management and implemented React-based frontend for user interaction
- Containerized the application with Docker and deployed via GitHub Actions CI/CD pipeline to AWS Elastic Beanstalk, integrating with Route 53 and Application Load Balancer for scalable, high-availability hosting
- Captured and stored structured conversation logs for downstream machine learning and statistical analysis to support ongoing research publications
- Applied advanced statistical modeling techniques such as regression analysis, time-series forecasting, and anomaly detection to identify trends in digital phenotyping datasets
- Built and automated SQL-based ETL workflows, reducing query execution time by 20% and improving data extraction efficiency for research teams
- Conducted A/B testing and statistical hypothesis testing to assess the impact of various health interventions and recommend data-driven strategies
- Developed interactive dashboards and data visualizations using Matplotlib, and Seaborn to communicate actionable insights improving research efficiency
- Collaborated with cross-functional teams (engineers, researchers, and healthcare experts) to translate complex data findings into strategic decisions, influencing decision-making and intervention strategies

Data Scientist Aug. 2019 – July 2022
Turnkey Learning *India*

- Engineered an NLP pipeline to automate adverse drug event reporting in alignment with pharmacovigilance standards and regulatory compliance requirements, reducing report processing time by 40%, utilizing Python, PyTorch, and AWS Lambda for scalable deployment
- Designed and implemented a machine learning model for user segmentation, risk scoring and classifying high-risk insurance claimants, enhancing ad targeting precision by 25% in large-scale datasets
- Improved scalability and maintenance of machine learning models by leveraging AWS Lambda for serverless execution, reducing infrastructure costs by 15%
- Reduced manual data entry errors using AWS Step Function to validate the population of essential fields in reports, leveraging AWS DynamoDB for data storage and retrieval
- Developed a React-based front-end for the Adverse Drug Event Report Pipeline and deployed it using AWS Amplify and GitHub, reducing deployment time and ensuring continuous delivery
- Utilized AWS S3 for secure data storage and AWS Lambda for scalable serverless processing of reports, ensuring high availability and minimal downtime

- Facilitated faster deployment with CI/CD pipelines using Docker and Github Workflows, streamlining the release process
- Applied hypothesis testing and A/B experimentation to optimize ML models
- Collaborated with cross-functional teams to deliver actionable insights, enabling data-driven decision-making, using data visualization tools like Matplotlib and Tableau

Projects

Customer Risk & Segmentation Model | *Python, AWS, Flask*

- Developed a predictive risk-scoring model to classify and segment users based on historical health record data, improving precision by 25%
- Designed and deployed a scalable ML pipeline using AWS Lambda and Flask-based REST APIs, reducing model inference latency and improving real-time decision-making
- Leveraged anomaly detection techniques to identify fraudulent or high-risk users, reducing false positives
- Optimized data preprocessing workflows using PySpark and SQL, reducing feature extraction time, improving efficiency in handling large datasets
- Designed data-driven insights dashboards using Matplotlib, helping stakeholders make informed decisions on risk-based targeting strategies

Knowledge Distillation using LLMs | *Python, PyTorch, Hugging Face Transformers*

- Leveraged Knowledge Distillation to train compact student models to demonstrate knowledge transfer and compression of information from LLMs
- Achieved student model's accuracy of 83.34% which is comparable to that of larger LLMs
- Analyzed the model performance in both online and offline learning modes, highlighting the benefits of incorporating student feedback into the teacher model's training

Image Segmentation Using Deep Learning | *Python, TensorFlow, AWS*

- Engineered and optimized the UNET model, achieving an IoU score of 0.73 on the validation set, significantly improving segmentation accuracy with a smaller number of trainable parameters
- Utilized Python, TensorFlow, and Keras for model development, alongside data augmentation techniques to enhance the model's robustness to varying conditions
- Deployed the model using AWS EC2 and integrated with AWS Lambda to generate segmented images using RESTful APIs
- Analyzed model performance and identified areas for improvement, such as tuning hyperparameters and refining the architecture, to enhance edge detection in distant objects

Education

University of Illinois at Chicago — Chicago, IL

Master of Science in Computer Science — *Aug. 2022 – May 2024*

Computer Vision, Natural Language Processing, Neural Networks, ML for Graphs, Data Mining & Text Mining

Certifications & Publications

Google Data Analytics Certificate - LINK
DeepLearning.AI TensorFlow Developer - LINK
Machine Learning: Stanford Online - LINK
Published Paper: "QWERTY Keyboard in Virtual Domain Using Image Processing" at the 2019 International Conference on Intelligent Computing and Control Systems (ICCS) - LINK